Proceedings of ICCM 2024

22nd International Conference on Cognitive Modelling[1]

Editor

Catherine Sibert

[1]Co-located with the 57th Annual Meeting of the Society for Mathematical Psychology

Preface

The International Conference on Cognitive Modelling (ICCM) is the premier conference for research on computational models and computation-based theories of human cognition. ICCM is a forum for presenting and discussing the complete spectrum of cognitive modelling approaches, including connectionism, symbolic modeling, dynamical systems, Bayesian modeling, and cognitive architectures. Research topics can range from low-level perception to high-level reasoning. In 2024, ICCM was jointly held with MathPsych – the annual meeting of the Society for Mathematical Psychology. The conference was held at the University of Tilburg from July 19th to July 22nd. An additional, virtual conference was held online from June 17th to June 21st. The proceedings from the in-person and the virtual conference are published jointly.

Acknowledgements

Papers in this volume may be cited as:
Lastname, A., Lastname, B., & Lastname, C. (2024). Title of the paper. In Sibert, C. (Ed.).
Proceedings of the 22nd International Conference on Cognitive Modelling (pp. fst_pg-lst_pg).
University Park, PA: Applied Cognitive Science Lab, Penn State.

Conference Committees

General and Program Chairs

Catherine Sibert	University of Groningen
Jelmer Borst	University of Groningen
Marco Ragni	TU Chemnitz

Review Committee

Amirreza Bagherzadehkhorasani	Penn State University
Francesca Borghesi	University of Turin
Laura van de Braak	Radboud University
Yiyang Chen	University of Kansas
Hillmer Chona	Penn State University
Michael Collins	Air Force Research Laboratory
Brendan Conway-Smith	Carleton University
Edward Cranford	Carnegie Mellon University
Taylor Curley	Air Force Research Laboratory
Chris Dancy	Penn State University
Maria Jose Ferreira	Carnegie Mellon University
Theodros Haile	University of Groningen
Lingyun He	Penn State University
Nils Wendel Heinrich	Universität zu Lübeck
Alexander R. Hough	Air Force Research Laboratory
Mary Kelly	Carleton University
Yung Han Khoe	Radboud University Nijmegen
Joshua Krause	University of Groningen
Othalia Larue	Parallax Advanced Research
Peter Lindes	University of Michigan
Leendert van Maanen	Utrecht University
Miki Matsumuro	Cornell University
Junya Morita	Shizuoka University
Kazum Nagashima	Shizuoka University
David Peebles	University of Huddersfield
Kai Preuss	Technische Universität Berlin
Stefan T. Radev	Heidelberg University
Hedderik van Rijn	University of Groningen
Frank Ritter	Penn State University
Serhii Serdiuk	Penn State University
Thomas Sievers	Universität zu Lübeck

Nisheeth Srivastava   Indian Institute of Technology Kanpur
Sönke Steffen   University of Groningen
Terry Stewart   National Research Council of Canada
Anastasia Stoops   University of Illinois at Urbana-Champaign
Elpida Tzafestas   National and Kapodistrian University of Athens

Table of Contents

Computer-Based Experiments in VR: A Virtual Reality Environment to Conduct Experiments, Collect Participants' Data and Cognitive Modeling in VR

Amir Bagherzadeh (amir.bagherzadeh@psu.edu)
Department of Industrial and Manufacturing Engineering, Penn State, University Park, PA 16802 USA
Farnaz Tehranchi (farnaz.tehranchi@psu.edu)
School of Engineering Design and Innovation, Penn State, University Park, PA 16802 USA

Abstract

In this paper, we explore the integration of Virtual Reality (VR) into behavioral experiments, addressing the technical challenges that researchers face due to the necessity of advanced programming and game engine knowledge. By developing a VR environment designed to provide an interface to move computer-based experiments to VR and pairing it with a *VR Analysis Tool* (VRAT) for data analysis and visualization, we facilitate a more accessible entry into VR-based research. The advantage that our tool provides is that researchers can transfer their traditional computer-based experiments to a VR environment with superior eye tracking and higher experiment validity due to a higher level of control over environmental factors. We also designed an experiment to compare VR eye-tracking systems with traditional screen-based eye-trackers in terms of accuracy and precision. We observed that based on our results, VR shows a better consistency across different screen sizes (24, 30 and 35-inch displays). Finally, we extended the capabilities of VisiTor, a tool that enables interaction for cognitive models to interact with the developed environment in VR.

Keywords: Behavioral studies; user study; virtual reality; eye tracking

Introduction

The evolution of human research experiments, particularly within the fields of psychology, cognitive science, and industrial engineering, has experienced a remarkable transformation over the past century. This journey from the simplicity of pen and paper to today's sophisticated, digitally driven methodologies signifies a significant leap in our capacity to understand human behavior and cognition. Initially, the field was characterized by using basic tools—paper, and pencils—which, while foundational, were constrained by their imprecision and the tedious nature of manual data collection and analysis. It is also susceptible to human error and subjective biases (Miller, 1956; Tolman, 1948). These primitive conditions placed inherent limitations on the scope and reliability of research findings, as the tools at hand could not fully capture or accurately measure the complexities of human cognitive processes and behaviors.

The use of computers enabled the presentation of stimuli and the recording of responses with greater accuracy and speed, supporting more complex experimental designs and sophisticated data analysis, thereby enriching the insights gleaned from such studies as decision-making studies in Choice Task environments (Newell et al., 2004), memory tests (Brunetti et al., 2014), etc.

Virtual Reality (VR) has been recognized as a potential tool in behavioral research for over two decades (Loomis et al., 1999). However, until recently, VR was limited to a small number of specialized labs due to technological limitations and the lack of accessible hardware and software. Advancements in technology have now made VR a viable tool for a wide range of behavioral researchers.

One of the most significant developments is the availability of powerful software engines, such as Unity (also known as Unity3D), which enable the creation of rich, immersive 3D environments. Unity, a popular game engine used for developing video games, animations, and other 3D applications, has seen a surge in popularity among researchers.

Unity offers well-developed systems for creating realistic graphics, simulating physics, and incorporating particles and animations. However, it lacks features specifically designed to cater to the needs of human behavior researchers. Recognizing this gap, we aimed to develop an open-source software resource that would allow researchers to harness the power of Unity for conducting behavioral studies.

Our goal is to provide a tool that empowers researchers to create and customize 3D environments tailored to their specific research questions, without requiring extensive programming knowledge. By leveraging the capabilities of Unity and integrating features relevant to behavioral research, we hope to facilitate the widespread adoption of VR in the field, leading to new insights and advancements in understanding human behavior.

The integration of VR technology into behavioral research has introduced a new level of immersion and control that previously was unattainable with conventional methods. VR's ability to simulate complex and realistic environments offers researchers a powerful tool to investigate human behavior with an unprecedented degree of validity. This technology enables the manipulation of environmental variables in controlled settings not imaginable before, allowing for the exploration of human behavior in contexts that are difficult, dangerous, or impossible to replicate in reality and modeling human reliability (Kheiri et al., 2023; Parsons & Rizzo, 2008; Slater, 2009).

Moreover, VR provides a novel platform for studying complex social interactions and cognitive processes, facilitating the examination of phenomena like empathy, social cognition, and decision-making within highly immersive, interactive environments (Blascovich et al., 2002; Fox et al., 2009).

The synergy between VR and eye-tracking technologies embodies a significant methodological enhancement in psychological research. Traditional and screen-based eye-tracking methods have provided invaluable insights into attention and cognition; however, they are often limited by the constraints of static environments, two-dimensional stimuli, the size of the display and the participants' distance. VR-based eye tracking rises above these limitations, offering a robust, dynamic, three-dimensional research environment that captures more naturalistic eye movements and behaviors.

In summary, the transition from pen-and-paper tests to advanced computer-based and now to VR-assisted experiments signifies a monumental shift in human research, enabling more nuanced, accurate, and impactful investigations. As these technologies continue to evolve, their integration promises to deepen our understanding of cognitive and behavioral phenomena, paving the way for innovative applications in psychology, education, and beyond.

In this paper, we present a series of tools that enable more researchers to conduct their studies in VR. We present the new environment we developed using Unity that enables researchers to easily conduct their current 2D-designed experiments in VR. This tool offers several advantages over traditional physical experimental setups.

Flexibility In Display Size

Researchers can adjust the display size and the distance to the screen to meet the specific requirements of their studies. The experiments with screen-based eye tracking are limited to the capabilities of the eye tracker. Conventional screen-based eye trackers have limitations when it comes to tracking users' gazes across multiple screens and are often cost-prohibitive for large displays (27-inch or larger). Although wearable eye trackers have been extensively utilized by researchers in this area (Kocejko et al., 2015; MacInnes et al., 2018) to overcome these challenges, they present their own set of issues.

One of the primary difficulties with wearable eye trackers is the complexity of accurately mapping eye gazes to display pixel coordinates. This process can be time-consuming and requires specialized software and calibration techniques. Additionally, wearable eye trackers often require special lenses for participants who wear eyeglasses, further increasing the overall cost of the research setup.

Motivated by these challenges, we sought to develop a more cost-effective and accurate method for eye tracking using VR technology. By leveraging VR, we aim to address two key issues:

1. Simplifying the process of mapping eye gazes into pixel coordinates, thereby improving the accuracy and reliability of the collected data.
2. Reducing the financial investment required for wearable eye trackers, making eye-tracking research more accessible to a wider range of researchers and institutions.

Our VR-based eye-tracking solution offers a promising alternative to traditional screen-based and wearable eye trackers. By harnessing the capabilities of VR technology, we can provide researchers with a more affordable, accurate, and user-friendly tool for conducting eye-tracking studies across multiple screens and large displays.

Enhanced Experiment Design

Our VR environment allows researchers to manipulate environmental variables in ways that are difficult to replicate in a real-world setting. VR technology and its derivatives are mature enough to be used in various settings, driving a surge in research across diverse professional environments like medicine, architecture, the military, and industry. This expansion of VR research brings to light the critical importance of meticulous experimental design, particularly the need for careful control of variables to ensure valid and reliable results. While VR offers researchers the ability to create highly immersive and interactive simulated environments, it also introduces a range of unique challenges in controlling potential confounding factors.

For instance, Bogacz et al. (2020) demonstrated that the choice of input device significantly impacts user behavior in a VR cycling simulation. Their study, comparing keyboard controls with an instrumented bicycle, revealed distinct differences in speed, acceleration, braking, and even head movements, highlighting the potential for input device choices to confound findings related to cognitive processing and decision-making (Bogacz et al., 2020). Additionally, (Chirico et al., 2018) explored the impact of different VR environments on the complex emotion of awe, emphasizing the importance of carefully designing control conditions to isolate the specific effects of the target stimuli. They found that even seemingly neutral VR environments can elicit emotional responses, underscoring the need for rigorous control over environmental variables to ensure accurate interpretations of results (Chirico et al., 2018).

Furthermore, research suggests that the level of user engagement in a VR task can influence their neural activity, particularly within the occipital alpha (α) frequency band (Klimesch et al., 1998). This finding highlights the importance of controlling for factors that might influence user engagement, such as task difficulty, instructions, and feedback, as variations in engagement could potentially confound results related to cognitive processes and emotional responses.
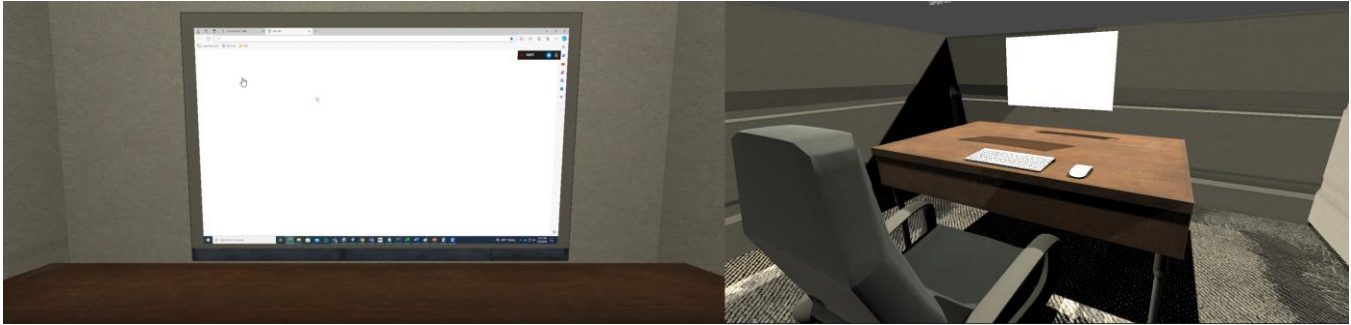
Figure 1: Participants point of view of the environment (left) and the environment design in Unity (right).

Also, the VR environment can be tailored quickly and easily to create the optimal conditions for multiple research purposes. And finally, our VR-based experimental setup is far more portable than setups reliant on physical displays, facilitating easier transportation and set up at various locations.

Improved Eye-Tracking

We show that eye-tracking within our VR environment can be more precise and reliable than traditional screen-based methods due to users' ability to move their heads freely and keep the fixation point at the center of the display. Research has shown that in immersive VR environments, users tend to move their heads to center interesting objects in their vision (Han et al., 2017; Han et al., 2019). This behavior is consistent with the "stare-in-the-crowd" effect, which is preserved in VR and can be influenced by factors such as social anxiety (Raimbaud et al., 2023). These findings suggest that users may keep their gaze point at the center of the screen in VR, particularly when interacting with virtual agents or exploring virtual environments.

This introduction of our VR tool highlights its potential to transform how research is conducted, offering a versatile and sophisticated platform for experimental design. This tool is designed to replicate traditional screen-based experiments in a simulated environment in VR.

Alongside the environment, we are also introducing an analysis tool called VR Analysis Tool (*VRAT)*, our new Python-based analysis tool, designed to work seamlessly with our VR platform. VRAT records on-screen activity, precisely tracking and saving gaze coordinates to generate detailed heatmaps of visual attention. After experiments, researchers can define their Areas of Interest (AOIs), and VRAT will label the collected gaze data that falls within these specified zones of the AOIs.

Additionally, we enhanced the capabilities of our previously developed tool, VisiTor, that enables cognitive models to interact with computer User Interfaces. With the latest enhancements, VisiTor now has the capability to control the VR headset and its controllers and interact with the developed environment.

The integrated application of these advanced tools revolutionizes the experimental design process, offering an unparalleled level of validity and accuracy in eye-tracking

data at a significantly reduced cost when compared to traditional methods. The affordability of VR headsets equipped with eye-tracking capabilities—like the Oculus Quest Pro or the HTC Vive Eye, priced at approximately $1,500—contrasts with the more expensive traditional screen-based eye trackers, which can exceed $4,000. Even the cheapest wearable eye trackers (e.g., pupil lab core) cost over $3700. Followed by employing VRAT for data analysis and leveraging VisiTor for developing accurate cognitive models, researchers can simulate and study human behavior with unprecedented fidelity and efficiency that was not possible before.

Methodology

VR Environment for Behavioral Research

The integration of computer technology in cognitive and behavioral research has significantly enhanced the precision and efficiency of experiments, particularly in areas like decision-making within choice-task environments (Bagherzadeh & Tehranchi, 2023). Additionally, the adoption of various input devices has made experiments more interactive and immersive, broadening the research scope and enhancing participant engagement. The transition to digital has also increased the reach and inclusivity of research, enabling the collection of data from diverse global populations through online platforms.

These methods have their shortcomings. Controlling all the conditions in physical environments is not easy. The setup is usually fixed, and it is not easy to replicate the environment in another location. Additionally, conventional screen-based eye trackers struggle with head movements (Hermens, 2015) and are limited by screen size constraints (eye trackers commonly support up to 27-inch (16:9 aspect ratio) displays and 60 cm distance between a user and the display). Moreover, data accuracy in screen-based eye trackers is significantly affected by direct sunlight or natural light exposure. Research by Hansen and Pece (2005) and Holmqvist and Andersson (2017) illustrates how infrared radiation from both sunlight and certain artificial lights can impair the efficacy of eye-tracking systems.

To address these challenges, we employ VR Headsets equipped with eye trackers, ensuring consistent environmental conditions, element placements, and

interactions for all participants. However, creating VR environments for behavioral experiments requires an advanced understanding of game engines and programming. Our solution involves an accessible environment for researchers, enabling them to execute studies within VR. We cast the computer's display content into a VR environment. The display is customizable, and the researcher can set the distance and the screen size to their desired measurements. The Participants' point of view and the Unity design window are shown in Figure 1.

We have implemented a pipeline, converting the 3D participants' eye gaze vectors into pixel coordinates on the screen if they are looking at the screen. Hence, in the pipeline, if the eye gaze vector hits the screen, we translate the location of the hit into precise pixel coordinates that are on the display. The gaze coordinates are saved in a CSV file, readily accessible for further examination by researchers.

VRAT for Analysis

Accurate eye gaze data are essential for meaningful analysis. However, without proper contextualization, these data are of little value. A critical aspect of analyzing eye movements is the integration of eye gaze data with display recordings. The conventional video recorder falls short as it fails to synchronize the eye gaze timestamps with corresponding video frames, risking inaccurate gaze-to-frame correlation and compromising the integrity of the analysis.
Our complementary analysis software, VRAT, integrates seamlessly with our environment. It will initiate the recording process simultaneously with the activation of the eye-tracking process. Hence, it ensures precise synchronization between eye-tracking data and video recording. Once an experiment concludes within the Unity platform, both eye-tracking and video recording are terminated immediately. Also, VRAT offers advanced features for detailed post-analysis, including the capability to overlay gaze points directly onto video frames of the content of the 2D display and generate a heat map video of the casted display. Areas of Interest (AOIs) can be defined with just a snapshot in VRAT and a column to the gaze data CSV file will be added that specifies the incidence of gaze points within AOIs (if any), alongside providing the functionality of creating a bounding box around the AOIs when they appear in the recording. Unlike other analysis tools that require manual AOI bounding in every frame, our tool uses advanced pattern-matching techniques and only requires a snapshot of the AOI which provides a significant advantage in tracking dynamic AOIs that might get resizes or move around during the eye-tracking session, thereby enhancing the analytical robustness and efficiency.

VisiTor (Vision + Motor): A Tool to Enable Interaction in VR

As cognitive modelers, our ultimate goal is to develop a cognitive model that replicates participants' behavior. The participants' movements in VR environments go beyond conventional means of interaction (such as mouse movements, clicks, keyboard inputs and attention shifts). VR introduces a complex layer of interactivity, where each head movement or interaction can lead to distinct visual scenes and states. Also, the means of interactions extend to VR Controllers and in some cases, tracking hands' movements.

As the first step, we extend the capabilities of our previously developed tool VisiTor (Bagherzadeh & Tehranchi, 2022) which enables cognitive models to interact with computer environments by providing mouse and keyboard inputs/outputs. VisiTor has already demonstrated its applicability in various contexts, such as simulating user behavior in a driving simulation game "Desert Bus for Hope" (Wu et al., 2023), modeling learning in a probabilistic game developed in HTML (Bagherzadehkhorasani & Tehranchi, 2023b), and mimicking motor impairments in typing tasks (Bagherzadehkhorasani & Tehranchi, 2023a).

With the new extensions, VisiTor can enable cognitive models to move the headset (changing position and head rotation), and key presses within the defined environments. VisiTor is powered by Python and communicates with the environment over TCP or Shell which makes it compatible with all cognitive architectures. This tool enables cognitive modelers to develop models that can interact with our virtual environment.

Through the seamless integration of our VR environment, VRAT, and VisiTor, we have established a pipeline that empowers cognitive modelers to conduct their conventional 2D experiments in VR, analyze user behavior with VRAT, and create interactive cognitive models with VisiTor.

VR vs. Screen-Based Eye-Tracking: Performance

In this section, we will propose an empirical one-to-one comparison experiment that can showcase differences in eye-tracking in VR compared to traditional screen-based methods across various screen sizes. The analysis can highlight the differences in accuracy and precision of VR-based eye tracking in comparison to traditional methods in different screen sizes and distances. We aim to provide an interface that can underscore the technological advancements and the improvements in participants' experience, thereby providing a compelling case for the adoption of VR technology in cognitive modeling and behavioral research.

Experiment Design

To assess the accuracy and precision of the eye trackers, we developed two distinct evaluation settings: one utilizing Python for 2D displays and another using Unity for VR. The evaluation protocols employed in both environments were designed to assure consistency for direct comparison. The procedure starts by calibrating the eye-tracking systems. Following calibration, the evaluation phase starts. The screen transitions to black, and a circle target measuring 10 pixels in radius, appears at the center (960,540-pixel coordinates). Participants are instructed to focus their gaze on the circle throughout its display duration. There is a one-second pause before gaze data collection begins, ensuring participants have
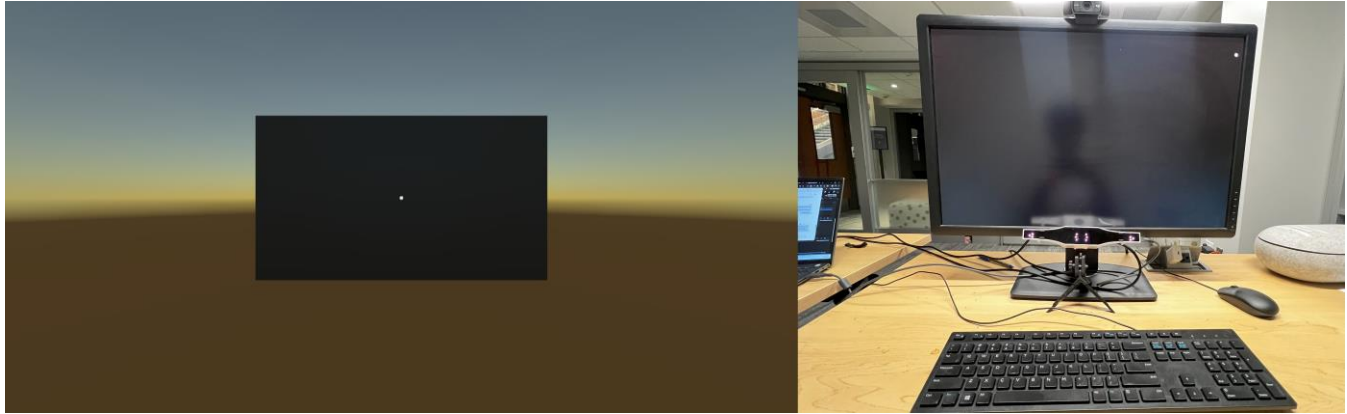
Figure 2: Participants' point of view of the accuracy and precision evaluation system in Virtual Reality (left) and Screen-Based system (right), the white dot is the target circle where users need to follow and keep their gaze on it.

adequate time to fixate on the target. Subsequently, gaze data are recorded for a two-second interval.

After this initial measurement, the target circle (radius = 20 pixels) relocates to various predefined coordinates sequentially: bottom right corner: (40,40), bottom left corner: (1880,40), top left corner (1880,1040), and finally top right corner: (40,1040), with the same data collection procedure, followed at each location. An illustrative Figure 2 shows the point of view of the participants in the evaluation environment within both the PC and VR settings for clarity.

For the computer-based environment, participants are positioned 70 cm from the monitor and advised to maintain their posture to limit movement-induced accuracy variances. Despite these instructions, slight posture adjustments are anticipated due to the natural discomfort associated with prolonged static positioning. In the VR setting, the eye-tracking system is integrated within the headset, allowing participants more freedom to move without jeopardizing data integrity. Nonetheless, it is crucial for the experimenter to ensure the headset is securely fitted to each participant's head to prevent slippages or movements that could affect the tracking precision and accuracy.

Procedure

For this study, three male grad students at Pennsylvania State University, aged between 25 and 30, participated in the evaluation experiment. They went through the evaluation process five times across three different screen sizes (24 inches, 30 inches, and 35 inches diagonally) using a screen-based eye tracking system. Subsequently, they underwent the same evaluation sequence within a VR environment, using virtual screens of equivalent sizes.

The screen-based eye tracking assessments utilized the GazePoint GP3 HD system (150Hz refresh rate, 0.5 – 1.0 degree of visual angle accuracy, Compatible with 24″ displays or smaller). It is one of the most used screen-based eye-tracking systems in research, while the Oculus Quest Pro, equipped with integrated eye tracking, served as the VR headset of choice. Unfortunately, Meta does not provide any information about their eye-tracking specifications. At the

time of writing this paper, GP3 HD costs $2,450.00 while Quest Pro is only $999.

Results

We conducted the screen-based tests using three different 1080p monitors corresponding to the specified sizes. It was observed that the screen-based eye tracker demonstrated inconsistent tracking capabilities, particularly with the largest screen size (i.e., 35-inch), leading to a significant data loss. Consequently, we deemed the accuracy and precision data for the 35-inch screen size as not applicable (NA). The results of the tests are summarized in Table 1 and Table 2.

Table 1 provides insights into the overall accuracy and precision of the screen-based eye-tracking system. Here, accuracy is defined as the pixel distance average Euclidean distance between the recorded gaze points and the intended focal point, while precision refers to the standard deviation of these distances.

The data suggests that measurement outcomes in screen-based systems vary significantly from user to user and this impact intensifies as screen sizes increase. Consequently, these findings suggest that VR-based eye tracking offers more consistent performance across users, which is particularly advantageous for research focused on eye movement phenomena like saccades and fixations, where uniform system performance is critical.

Table 1: Overall accuracy and precision of different methods of eye tracking across participants for screen size 30-inch in VR and Computer-Based (CB) eye tracking systems.

Participants	method	Overall Accuracy	Overall Precision
P 1	VR	88.5	14.8
	CB	71.2	47.6
P 2	VR	95.6	12.0
	CB	1058.3	258.7
P 3	VR	73.6	14.15
	CB	229.5	40.23

Table 2: The average accuracy and precision (across participants and points) for different screen sizes in VR and Computer-Based (CB) eye-tracking systems.

Method	Screen size	Across participants accuracy	Across participants precision
VR	24	144.8	32.4
CB	24	94.2	41.8
VR	30	85.9	13.7
CB	30	462.62	115.15
VR	35	101.1	43.7
CB	35	NA	NA

Table 2 suggests that VR eye tracking seems to be more consistent across casted various screen sizes in VR, underscoring a key advantage over traditional screen-based systems. As screen size grows, traditional eye trackers exhibit declining accuracy, struggling notably with very large displays (35 inches and above). In contrast, VR eye-tracking systems maintain their tracking efficacy irrespective of screen size. This advantage stems from the ability of VR users to move their heads, positioning target points centrally within their field of vision more naturally (Khan & Lee, 2019), as opposed to the strain of eye movements towards screen extremities required with large physical monitors in our study, the average angle between the user's eye gazes and the center of their vision was measured to be 11 degrees for evaluation point at the corners of the casted screens (almost indifference to screen size) while this number for our screen-based system was 23 for 24-inch, 28 for 30-inch and 32 for 35-inch monitors. Thus, based on our limited observation, VR not only enhances eye-tracking accuracy but also alleviates the physical strain on participants, offering improved user experience and data reliability.

Conclusion and Future Works

This paper is trying to bridge the gap between behavioral experiment design and VR. Designing studies in VR requires game-engine knowledge and advanced programming skills, which discourage researchers from conducting studies in VR. Hence, we designed an environment in VR in which researchers can conduct their computer-based experiments in VR. We integrated our environment with VRAT, an analysis tool capable of generating insightful visualizations and quantifiable outputs along with extending VisiTor to enable cognitive researchers to develop models that simulate users' behavior in VR.

Our study suggests that VR eye-tracking systems offer enhanced accuracy and precision over traditional screen-based eye trackers, particularly as screen size expands. However, it is important to note that our sample size was small, which limits the generalizability of our findings. This preliminary evidence indicates a potential trend where VR technology could provide more consistent and reliable measurements of eye movements across various screen sizes and users.

Through a comparative analysis between screen-based and VR environments, we demonstrated that VR eye tracking alleviates several limitations inherent in conventional methods. Specifically, the ability for participants to move their heads within a VR environment, maintaining target points at the center of their vision, significantly enhances tracking accuracy and user comfort, particularly for larger virtual screen sizes.

As we look to the future, the potential applications of VR-based eye tracking in cognitive science, psychology, user experience research, seem to be promising. By embracing these advanced technologies, researchers can achieve more insights into human visual attention and cognition, ultimately unlocking new possibilities for understanding and interfacing with human behavior. However, VR environments have their own shortcomings, motion sickness is the most known issue with VR headsets. Specially, through long sessions. However, with the ever-increasing interest in VR, more research (Becker & Ngo, 2016; Curtis et al., 2015; Nie et al., 2019) is being done to tackle this issue.

Due to the applicability of the 3D design, the next generation of behavioral studies will be conducted in VR and Cognitive architectures should support and simulate different means of interactions. VisiTor is the first step toward this path. This tool is integrated with our environment and will provide a way to translate Cognitive models' interaction into actionable movement in VR. In the near future, we plan to publish VisiTor controller package, which will enable cognitive models to interact with VR environments operatable by SteamVR.

References

Bagherzadeh, A., & Tehranchi, F. (2022). Comparing cognitive, cognitive instance-based, and reinforcement learning models in an interactive task. Proceedings of ICCM-2022-20th International Conference on Cognitive Modeling,

Bagherzadeh, A., & Tehranchi, F. (2023). The Analysis of the Effect of Visual Cues in a Binary Decision-Making Environment. Conference proceedings AHFE,

Bagherzadehkhorasani, A., & Tehranchi, F. (2023a). Automatic Error Model (AEM) for User Interface Design: A new approach to Include Errors and Error Corrections in a Cognitive User Model. Proceedings of the Human Factors and Ergonomics Society Annual Meeting,

Bagherzadehkhorasani, A., & Tehranchi, F. (2023b). *A Pipeline for Analyzing Decision-Making Processes in a Binary Choice Task* Proceedings of the 21th International Conference on Cognitive Modelling, Amsterdam, Netherland.

Becker, J., & Ngo, T. (2016). Mitigating visually-induced motion sickness in virtual reality. *Unpublished manuscript. Stanford University, Stanford, CA*.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological inquiry*, *13*(2), 103-124.

Bogacz, M., Hess, S., Calastri, C., Choudhury, C. F., Erath, A., van Eggermond, M. A., Mushtaq, F., Nazemi, M., & Awais, M. (2020). Comparison of cycling behavior between keyboard-controlled and instrumented bicycle experiments in virtual reality. *Transportation research record*, *2674*(7), 244-257.

Brunetti, R., Del Gatto, C., & Delogu, F. (2014). eCorsi: implementation and testing of the Corsi block-tapping task for digital tablets. *Frontiers in psychology*, *5*, 939.

Chirico, A., Ferrise, F., Cordella, L., & Gaggioli, A. (2018). Designing awe in virtual reality: An experimental study. *Frontiers in psychology*, *8*, 293522.

Curtis, M. K., Dawson, K., Jackson, K., Litwin, L., Meusel, C., Dorneich, M. C., Gilbert, S. B., Kelly, J., Stone, R., & Winer, E. (2015). Mitigating visually induced motion sickness: a virtual hand-eye coordination task. Proceedings of the Human Factors and Ergonomics Society Annual Meeting,

Fox, J., Bailenson, J., & Binney, J. (2009). Virtual experiences, physical behaviors: The effect of presence on imitation of an eating avatar. *Presence: Teleoperators and Virtual Environments*, *18*(4), 294-303.

Han, H., Lu, A., & Wells, U. (2017). Under the movement of head: evaluating visual attention in immersive virtual reality environment. 2017 International Conference on Virtual Reality and Visualization (ICVRV),

Han, H., Lu, A., Xu, C., & Wells, U. (2019). Object-based Visual Attention Quantification using Head Orientation in VR Applications. *International Journal of Performability Engineering*, *15*(3), 732.

Hansen, D. W., & Pece, A. E. (2005). Eye tracking in the wild. *Computer Vision and Image Understanding*, *98*(1), 155-181.

Hermens, F. (2015). Dummy eye measurements of microsaccades: Testing the influence of system noise and head movements on microsaccade detection in a popular video-based eye tracker. *Journal of Eye Movement Research*, *8*(1).

Holmqvist, K., & Andersson, R. (2017). Eye tracking: A comprehensive guide to methods. *Paradigms and measures*.

Khan, M. Q., & Lee, S. (2019). Gaze and eye tracking: Techniques and applications in ADAS. *Sensors*, *19*(24), 5540.

Kheiri, S. K., Vahedi, Z., Sun, H., Megahed, F. M., & Cavuoto, L. A. (2023). Human reliability modeling in occupational environments toward a safe and productive operator 4.0. *International Journal of Industrial Ergonomics*, *97*, 103479.

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., & Schwaiger, J. (1998). Induced alpha band power changes in the human EEG and attention. *Neuroscience letters*, *244*(2), 73-76.

Kocejko, T., Ruminski, J., Wtorek, J., & Martin, B. (2015). Eye tracking within near-to-eye display. 2015 8th International Conference on Human System Interaction (HSI),

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior research methods, instruments, & computers*, *31*(4), 557-564.

MacInnes, J. J., Iqbal, S., Pearson, J., & Johnson, E. N. (2018). Wearable Eye-tracking for Research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices. *BioRxiv*, 299925.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Newell, B. R., Rakow, T., Weston, N. J., & Shanks, D. R. (2004). Search strategies in decision making: The success of "success". *Journal of Behavioral Decision Making*, *17*(2), 117-137.

Nie, G.-Y., Duh, H. B.-L., Liu, Y., & Wang, Y. (2019). Analysis on mitigation of visually induced motion sickness by applying dynamical blurring on a user's retina. *IEEE transactions on visualization and computer graphics*, *26*(8), 2535-2545.

Parsons, T. D., & Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of behavior therapy and experimental psychiatry*, *39*(3), 250-261.

Raimbaud, P., Jovane, A., Zibrek, K., Pacchierotti, C., Christie, M., Hoyet, L., Pettré, J., & Olivier, A.-H. (2023). The Stare-in-the-Crowd Effect When Navigating a Crowd in Virtual Reality. ACM Symposium on Applied Perception 2023,

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3549-3557.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Wu, S., Bagherzadeh, A., Ritter, F., & Tehranchi, F. (2023). Long road ahead: Lessons learned from the (soon to be) longest-running cognitive model. 21st International Conference on Cognitive Modeling (ICCM) at the University of Amsterdam, the Netherlands,

Genetically Evolving Verbal Learner:
A Computational Model Based on Chunking and Evolution

Dmitry Bennett* (D.Bennett5@lse.ac.uk)
Centre for Philosophy of Natural and Social Science, Houghton Street,
London School of Economics, WC2A 2AE, United Kingdom

Noman Javed* (N.Javed3@lse.ac.uk)
Centre for Philosophy of Natural and Social Science, Houghton Street,
London School of Economics, WC2A 2AE, United Kingdom

Laura Bartlett (L.Bartlett@lse.ac.uk)
Centre for Philosophy of Natural and Social Science, Houghton Street,
London School of Economics, WC2A 2AE, United Kingdom

Peter Lane (p.c.lane@herts.ac.uk)
Department of Computer Science, College Lane,
University of Hertfordshire, AL10 9AB, United Kingdom

Fernand Gobet (F.Gobet@lse.ac.uk)
Centre for Philosophy of Natural and Social Science, Houghton Street,
London School of Economics, WC2A 2AE, United Kingdom

Abstract

A fundamental issue in cognitive science concerns the interaction of the cognitive "how" operations, the genetic/memetic "why" processes, and by what means this interaction results in constrained variability and individual differences. This study proposes a single GEVL model that combines complex cognitive mechanisms with a genetic programming approach. The model evolves populations of cognitive agents, with each agent learning by chunking and incorporating LTM and STM stores, as well as attention. The model simulates two different verbal learning tasks: one that investigates the effect of stimulus-response (S-R) similarity on the learning rate; and the other, that examines how the learning time is affected by the change in stimuli presentation times. GEVL's results are compared to both human data and EPAM – a different verbal learning model that utilises hand-crafted task-specific strategies. The semi-automatically evolved GEVL strategies produced good fit to the human data in both studies, improving on EPAM's scores by as much as factor of two on some of the pattern similarity conditions. These findings offer further support to the mechanisms proposed by chunking theory, connect them to the evolutionary approach, and make further inroads towards a Unified Theory of Cognition (Newell, 1990).

Keywords: learning; chunking; genetic programming; GEMS; CHREST; learning; long-term memory; short-term memory.

Introduction

Understanding intelligent organisms requires answering two questions. The first of these is "How?" The "how" question is what the brain is tasked to solve: *how* to navigate by stars, *how* to weave a spider web, *how* to learn to recognise a predator or a danger, *how* to learn in general? In order to solve the numerous "how" questions, the brain has to process sensory input data, recognise objects, store, update and retrieve memories – it must form internal representations of the world and utilise them in its computations. In psychology, understanding of the "how" question is tackled both by cognitive-based models (which unravel the mechanisms that underlie cognition, starting with high-level cognition and behaviour levels) and neuroscience-based models (which focus on the lower-level neural functions).

The other fundamental question is "Why?" For example, *why* does an organism navigate by stars or weave a spider web? The answer is: it does so to spread the replicators that brought about this organism and its behaviour (not to be confused with merely spreading the said organism's offspring). The replicators may include the purely genetic kind (i.e., the DNA), but, in some cases, also the informational/cultural type (i.e., "memes") (Dawkins, 1976, 1982; Hunter, 2018; Ridley, 2016).

In fact, the two questions are intertwined, with an organism's "how" strategies being influenced both by its more basic level "how" mechanisms and by the "why" selection pressure.

It is typical for psychological models to focus on the "how" question (see Kotseruba and Tsotsos (2020) for a review of over 40 recent cognitive models). These models produce complex simulations of interacting cognitive or neural structures (such as the LTM, STM, selective attention, chunks, or, synapses and networks of neurons). The functions of the models are hand-tuned for task specific behaviour (e.g., Lieto, 2019; Richman, Simon, & Feigenbaum, 2002).

---

* Both authors contributed equally

Other models start with the "why" question. They use genetic algorithms or genetic programming as a way to search and optimise agents' cognitive "how" strategy space (e.g., Brave, 1996; Lane et al., 2014). This implies that the task-specific behaviour is not hand-tuned by the model creators, but is semi-automatic (the initial program state is hand-crafted, but the subsequent mutation, generation and selection of programs is fully automatic). However, while the resultant program/strategy set of an evolved agent's may be long and complex, the inherent cognitive structures are often rudimentary – e.g., lacking simulations of LTM and perceptual mechanisms (for example, see Bartlett et al., 2023; Gunaratne & Patton, 2022).

The aim of this paper is to bridge the "Why?" and the "How?" approaches by integrating a complex cognitive-based psychological model and a genetic programming system into a single whole. This makes it possible to keep the inherent complexity of the psychological structures and mechanisms, while adding the automaticity of task-specific strategy discovery. To achieve this aim, we will combine the CHREST cognitive architecture with GEMS, an environment for semi-automatically evolving cognitive models.

The Cognitive-Based CHREST Model

CHREST (Chunking Hierarchy and REtrieval STructures) (Gobet, 1993, 2000; Gobet & Lane, 2012; Gobet & Simon, 2000) is a formal cognitive model based on one of the most established theories in cognitive psychology – the chunking theory (Chase & Simon, 1973; Gobet et al., 2001; Simon, 1974).

The core concept of the chunking theory – a *chunk* – can be defined as a meaningful unit of information constructed from elements that have strong associations between each other (e.g., a group of digits making up a phone number). Thus, *chunking* is the process of forming and updating chunks in the LTM (Gobet, Lloyd-Kelly, & Lane, 2016; Simon, 1974). Although the chunks themselves vary between people due to personal differences, chunking mechanisms are largely invariant across domains, individuals, and cultures (Chase & Simon, 1973; Gobet et al., 2001; Miller, 1956; Simon, 1974).

Beyond verbal descriptions, chunking mechanisms have been formalised into computer programs – first EPAM (Feigenbaum and Simon, 1962) and now CHREST (Chunking Hierarchy REtrieval STructures) (Gobet, 1993, 2000; Gobet & Lane, 2012; Gobet & Simon, 2000).

CHREST is an idealised self-organising cognitive system that simulates human learning. Chunks are operationalized as nodes in a graph and chunking is the process of adding new data to the LTM. This is done via two psychologically plausible cognitive processes: discrimination and familiarisation. Discrimination is the process of adding a new node to the LTM network. Familiarisation updates existing nodes with new information. Thus, learning is influenced both by the environmental stimuli and the data that have already been stored (Gobet & Lane, 2012). CHREST's STM

structure allows for additional ways to create links between chunks, such as linking chunks across visual and verbal modalities, or linking stimuli and responses within a single modality.

Moreover, chunking theory postulates time costs for its core cognitive operations, with discrimination taking ten seconds, familiarisation taking two seconds, and recognition of a pattern requiring around one hundred milliseconds. More generally, CHREST's learning may be described by a power law function, with repetitions of a task leading to diminishing improvements in performance accuracy and speed (with some caveats – e.g., single-shot learning being also possible, depending on the presented stimulus and chunks already stored in the LTM).

EPAM, CHREST and their derivative models were used to predict and postdict behaviour in verbal learning research (Feigenbaum, 1959; Feigenbaum & Simon, 1984; Richman & Simon, 1989; Richman et al., 2002) and accounted for context effects in perception, various aspects of concept formation (Bennett, Gobet, & Lane, 2020; Lane & Gobet, 2012), problem solving (Lane, Cheng, & Gobet, 2000), acquisition of syntactic categories (Freudenthal et al., 2016), emotion processing in problem gambling (Schiller & Gobet, 2014), developmental trends and cognitive decline due to ageing (Mathy et al., 2016; Smith, Gobet, & Lane, 2007), expert behaviour (Gobet & Simon, 2000; Richman et al., 1996; Richman, Staszewski, & Simon, 1995; Simon & Gilmartin, 1973), and the list goes on.

For further details of the chunking theory and CHREST see Gobet and Lane (2012).

The Genetic Side – GEMS

GEMS (Genetically Evolving Models in Science) is a modelling framework that is used to create cognitive models (Bartlett et al., 2023; Lane & Gobet, 2013). This system integrates insights from computational cognitive science, experimental psychology, and evolutionary computation to produce cognitive models in the form of computer programs. These models undergo evaluation on a simple cognitive architecture that mimics the execution of cognitive processes in the brain. The basic architecture of GEMS is presented in Figure 1.
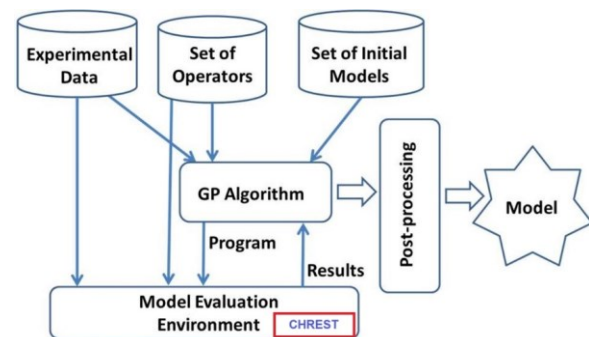

Figure 1. Overview of GEMS and its interaction with CHREST.

GEMS relies on Genetic Programming (GP) to explore the search space of candidate cognitive models. Initially, a predetermined number of candidate models are generated randomly. Subsequently, the GP algorithm iteratively combines and modifies these models using processes like mutation and crossover, guided by fitness evaluations based on the fit between the predictions of the models and the experimental data. By evolving models iteratively across multiple generations, GP ultimately proposes several typically successful solutions to the problem under consideration.

GP depends on a set of operators, serving as fundamental building blocks essential for constructing cognitive models within our meta-modelling system. These operators, akin to basic functions or operations implemented as programming functions, play a crucial role in accurately simulating cognitive processes by processing information across the components of the cognitive architecture. The selection of these operators is carried out by domain experts, carefully assigning relevant semantics and timings based on the relevant research literature. The choice of operators is also contingent on the experiment being simulated; for instance, for a short-term memory task, long-term memory operators may not be necessary.

GP generates multiple programs by combining these operators. A model evaluation environment is necessary to rate the quality of these programs, known as fitness evaluation in the language of GP. Our system diverges from most other GP systems at this point. We simulate experimental conditions and provide real data to these models. Consequently, these models act like humans undergoing psychological experiments under tightly controlled laboratory conditions. They undergo the same experimentation, and their responses, response times, and other relevant variables are measured. These measurements are then compared with those of human subjects. Importantly, our objective is not to evolve highly efficient models but rather to evolve models that closely simulate human behaviour. Therefore, the closer the resemblance to human data, the higher the fitness of the model.

At the end of a run of the GP system, it yields a population of good quality solutions. We have incorporated several post-simplification and analysis functionalities into the system to further simplify these models and cluster them based on their similarities. However, all this is contingent on the requirements of the experiment one is trying to simulate.
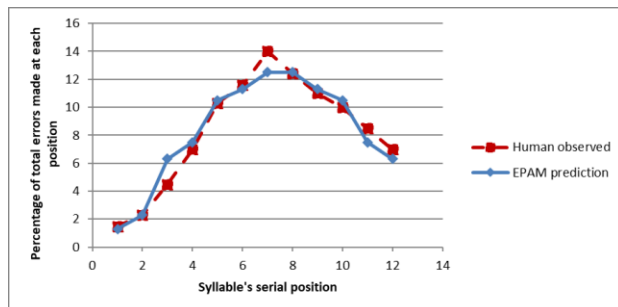
Figure 2. EPAM simulation of primacy-recency serial position curve. Adapted from (Feigenbaum & Simon, 1962)

Verbal learning

Verbal learning involves teaching participants lists of paired stimulus-response (S-R) nonsense syllables to uncover the fundamental laws of learning. Famous examples of verbal learning research include the "magic number 7 (plus or minus two)" study that established the STM capacity to be around seven chunks (Miller, 1956), and the study of the primacy-recency effect (where people were found to make fewest mistakes at the beginning and at the end of a memorised sequence) (McCrary & Hunter, 1953). Another, less known, but no less significant contribution of the verbal learning research, was how it was used to develop and shape the mechanisms of cognitive models such as EPAM and CHREST. Indeed, Richman et al. (2002) reported that various versions of EPAM have captured at least 20 regularities that were picked up by research into human rote learning. For example, the "intralist similarity" effects (Hintzman, 1968, 1969) were successfully simulated by EPAM: humans (and EPAM) produced more errors in S-R learning trials when the stimuli – nonsense consonant trigrams – were similar to each other (e.g. "ZIK" and "ZYJ") than when they were dissimilar. It also successfully simulated forgetting, in the shape of "oscillations" that occur during the learning of a single list and in the shape of "retroactive inhibition" that happens when learning a second list disrupts memory of a first list (Feigenbaum & Simon, 1962; Thune & Underwood, 1943).

EPAM also accounted for the primacy-recency serial position curve that describes people's tendency to remember first and last items better than middle-of-the-list ones (Feigenbaum, 1963; McCrary & Hunter, 1953) (see Figure 2). Finally, single-trial learning (Rock, 1957) was also explained and postdicted by EPAM (Gregg & Simon, 1967) – it was dependant on the complexity/simplicity (as determined by EPAM's LTM network) of the stimuli and upon the attention strategy of the participant.

EPAM's cognitive time cost parameters explained humans' constant learning time that was independent of the number of stimuli presentations (Bugelski, 1962; Richman et al., 2002).

The above research was important for establishing rigorously defined links between major cognitive structures and learning. However, one of its shortcomings was its reliance on learning strategies that were handcrafted by psychology experts. For example, EPAM's (and CHREST – which inherited most of EPAM's mechanisms) default S-R learning strategy was pre-defined as "learn stimulus as little as possible, before switching attention to the response; learn the response fully" (e.g., only learn letter "B" from the "BAJ" stimulus, but learn the response "WOJ" fully). Despite the hand-crafting, the models' fit to human data was often lacking; for example, the number of learning trials between the EPAM and the human data differed by as much as a factor of two (see Table 3).

The Present Study

The current study aims to replicate the verbal learning tasks reported by Underwood (1953), Bugelski (1962) and

simulated by EPAM (Richman et al., 2002). Crucially, the bridging of CHREST and GEMS approaches will allow us not only to integrate the "how" and the "why" aspects of a cognitive system, but also to move away from hand-crafting task-specific attentional and learning strategies, and to optimise the fit of the models to the human data.

Table 1. S-R syllables used in the intralist similarity task.

		SIMILARITY
Stimuli	**Low**	XIL, TOQ, WEP, DUF, MIZ, JUK, NAS, HOV, BIR, GAC
	Medium	HIZ, VEC, VIR, JUW, HUL, FEC, YOR, JAL, FOZ, YAW
	High	HUX, HEX, YAL, YOR, JIR, YOL, JAX, JIX, JER, HUL
Responses	**Low**	VOD, HAX, CEM, KIR, SIQ, FEP, BAJ, LOZ, TUW, YUG
	Medium	HIZ, VEC, VIR, JUW, HUL, FEC, YOR, JAL, FOZ, YAW
	High	HUX, HEX, YAL, YOR, JIR, YOL, JAX, JIX, JER, HUL

Method

Training and Testing

The human data for this project were taken from previous research: the investigation of the effect of pattern similarity on learning rates (Underwood, 1953), and the study into the effect of stimuli presentation time on the learning rate (Bugelski, 1962). The EPAM data were taken from Richman et al. (2002).

The stimulus-response patterns were as follows. There were 3 × 10 S-R pairs: Low-Low, Medium-Medium, and High-High. The low similarity stimuli and responses contained 20 different consonants, the medium similarity

recognise stimulus, recognise and learn stimulus, recognise and learn response, learn the link between the stimulus and the response, repeat operation, and wait for one or two seconds (see Table 2). The models were trained on "per condition" basis (see below), with population size 500 and 50 generations; the mutation rate was set to 0.2 (the latter was empirically selected to improve diversity in the population of strategies); there were independent runs for all conditions. The fitness was a sum of recall errors and the absolute difference between the number of trials taken by humans and GEVL.

We replicated the verbal learning experiments on pattern similarity and constant learning time, by Underwood (1953) and Bugelski (1962) respectively. Underwood's pattern similarity experiment involved five conditions: Low-Low, Low-Medium, Low-High, Medium-Low, High-Low. Each condition had a set of corresponding ten S-R pairs. Just like the human participants in the original studies, the models were presented with a stimulus (from one of the S-R pairs) for two seconds. If the models did not provide the correct response, they would be presented with the stimulus and response for four seconds. This trial cycle would repeat until the models provided correct responses for each of the ten stimuli.

The constant learning time simulations replicated Bugelski's experiment with human participants. There were five experimental conditions that varied the stimulus presentation time: 6 sec, 8 sec, 10 sec, 12 sec, and 19 sec. The models were presented with a list of ten S-R pairs, one pair at

Table 2. Overview of GEVL operators. Each operator type had a time cost (in milliseconds, ms) as follows: input (100 ms), output (140 ms), LTM (2000 ms for familiarisation, or 10000 ms for discrimination), syntax (0 ms).

Operator	Function	Type
PROG-X	a sequence of 2, 3 or 4 subprograms	Syntax
REPEAT2	repeats a subprogram 2 times	Syntax
ATTEND-STIMULUS	place the stimulus value into input slot 1	Input
ATTEND-RESPONSE	place the response value into input slot 2	Input
REC-AND-LEARN-ST	calls CHREST's recognise-and-learn-pattern function to learn a stimulus	LTM
REC-AND-LEARN-RES	calls CHREST's recognise-and-learn-pattern function to learn a response	LTM
RECOGNISE-ST	calls CHREST's recognize-pattern function to locate a pattern in long-term memory	LTM
LEARN-AND-LINK	calls CHREST's learn-and-link-two-patterns function to associate stimulus with response	LTM
RESPOND	retrieve the linked pattern using the stimulus and assign it to the model's output slot	Output
WAIT-X	advances model-clock (in ms): 1000 or 2000	Time

stimuli and responses contained 10 different consonants, and high similarity stimuli and responses contained 6 different consonants (see Table 1).

Procedure

In order to automatically generate verbal learning strategies, CHREST was integrated with GEMS to produce GEVL – Genetically Evolving Verbal Learner. Every trial consisted of a population of evolved models being presented with a S-R pair, and selecting a list of verbal-learning operators. The operators included attend stimulus, attend response,

a time, with each pair being presented for the duration that corresponded to the experimental condition. The experimental trial was repeated until the models learnt correct responses for all the stimuli.

Results

The results of the best GEVL verbal learning models are presented in Table 3 and Table 4. GEVL was able to achieve good fit to human data in Underwood's pattern similarity study. A sample strategy for the Low-High condition is

presented in Figure 3. GEVL achieved a near constant presentation time to trials ratio in its simulation of the Bugelski task. A sample strategy for the Bugelski task is in Figure 4. The r^2 is 0.99 for both experiments; for the Underwood task, RMSE was 0.37; for the Bugleski task, RMSE was 0.24.

As a type of sensitivity analysis, we used GEVL with mismatching S-R pairs and conditions (e.g., optimising for High-Low human result, when using Low-Low stimuli). The resulting fit was poor (the models never reached 30 trials, but topped out at 23 trials). This shows that the psychological constraints embodied in GEVL impose limits on its ability to fit the "human data". This also shows that the models did not merely resort to padding strategies with time delaying operations, but discovered important learning strategies.

We should note that we are reporting only the results with the best models, with there being many models that achieved similarly high fit.

Please see https://github.com/Voskod/GEVL for the code and the best models for all the experimental conditions.

```
(PROG4
    (PROG4 (LEARN-AND-LINK)
        (PROG4 (RECOGNISE-ST) (ATTEND-STIMULUS)
            (PROG3 (RESPOND) (ATTEND-RESPONSE)
            (ATTEND-STIMULUS))
        (REC-AND-LEARN-ST))
        (REC-AND-LEARN-ST) (RESPOND))
        (PROG4 (RESPOND) (ATTEND-RESPONSE) (LEARN-
AND-LINK)
        (ATTEND-STIMULUS))
        (WAIT-1000) (REPEAT2 (ATTEND-RESPONSE)
(ATTEND-RESPONSE)))
```

Figure 3. One of the strategies in the final population of models that learnt S-R pairs in the Low-High pattern similarity condition (Underwood's experiment).

```
(PROG4
    (REPEAT2
        (REPEAT2 (REC-AND-LEARN-RES)
        (REPEAT2 (ATTEND-RESPONSE) (RESPOND)))
        (PROG2 (ATTEND-STIMULUS) (REC-AND-LEARN-
RES)))
    (REPEAT2
        (PROG3 (LEARN-AND-LINK)
        (REPEAT2 (REC-AND-LEARN-RES) (WAIT-1000))
        (PROG3 (WAIT-2000) (RECOGNISE-ST) (REC-AND-
LEARN-RES))
        (LEARN-AND-LINK))
        (PROG3 (REC-AND-LEARN-RES) (REC-AND-LEARN-
RES)
        (REPEAT2 (REC-AND-LEARN-RES) (WAIT-1000)))
        (REPEAT2
        (REPEAT2 (WAIT-2000)
        (PROG3 (LEARN-AND-LINK) (WAIT-2000) (LEARN-
AND-LINK)))
        (PROG2
        (PROG3 (RECOGNISE-ST) (REC-AND-LEARN-RES)
        (LEARN-AND-LINK))
        (REC-AND-LEARN-RES))))
```

Figure 4. One of the strategies for Bugelski's constant learning experiment, for the 19-second condition.

Table 3. The effect of the S-R pattern similarity on the number of learning trials in humans, EPAM VI and GEVL. Human data are from Underwood (1953), EPAM VI data are from Richman et al. (2002).

Condition	People	EPAM VI	GEVL
Low-Low	23.2	13.4	23.0
Low-Medium	22.4	13.0	22.0
Low-High	24.4	13.0	24.0
Medium-Low	25.5	15.3	26.0
High-Low	30.7	16.0	31.0

Table 4. The effect of S-R presentation time on the number of learning trials in humans, as well as EPAM VI and GEVL simulations. Human data are from Bugelski (1962), EPAM VI data are from Richman et al. (2002).

Presentation Time	People	EPAM VI	GEVL
6 sec	10.2	9.3	10.0
8 sec	8.8	7.2	9.0
10 sec	5.8	5.9	6.0
12 sec	4.7	5.0	5.0
19 sec	3.3	3.5	3.0

Discussion

There are several key strengths and contributions of the current study. First, our approach integrated a cognitive model (with complex simulations of the LTM and STM) and a genetic programming environment, thus allowing to capture individual differences in learning. Our model demonstrated that there may be multiple solutions that satisfy a particular set of constraints. This is fully in line with research on individual differences in psychology – there is no *one* cognitive system in nature, there is inherent variability. Of course, this variability is constrained. For example, individual bees vary in their social behaviour, as do humans, but the intraspecies variability is bounded by species-specific physiological and cognitive structures (Crespi, 2014, 2017; Rubenstein & Hofmann, 2015). In our case, the evolved agents shared the basic cognitive mechanisms and structures (as operationalised by CHREST), but differed in their approaches to S-R learning. For example, one model in the final population had a S-R learning strategy that contained 22 cognitive operations, while another model contained 31 operations. This study is a rigorous demonstration of how the informational environment may shape both the cognitive strategies and the population of cognitive agents.

Secondly, our GEVL model moved away from hand-crafted learning strategies that were used in previous verbal learning research. Indeed, while EPAM prescribed rigid attentional shifts in S-R tasks (Richman et al., 2002), our model developed a wide range of strategies. For example, while EPAM was preconfigured to always learn just the first letter of the "Low" stimulus before learning the "High" response fully. On the other hand, GEVL's "Low-High" strategies were much more varied, with attention oscillating between stimuli and response multiple times.

Third, GEVL did a good job with simulating human data with regard to the number of trials needed to learn the patterns. For example, humans learn the "High-Low" set of S-R pairs in around 31 trials, while EPAM model takes approximately 16 trials – despite its hand-crafted verbal learning task-specific strategies. This is in contrast to GEVL, which produces a group of strategies that take around 31 trials (in general, for the pattern similarity task, EPAM's RMSE was 11.26 and r^2 was 0.77, versus GEVL's 0.37 and 0.99 respectively). With that said, it is important not to overstate the automatic nature of GEVL as the initial state of our model is also hand-crafted (i.e., the CHREST cognitive architecture, the GEMS environment for evolving models, and the choice of cognitive operators).

Another strength of GEVL is that the model is not a "black box," but is readily interpretable – both in terms of its underlying structures and the produced sets of cognitive strategies. On a related note, it is an interesting research question if the current study could be replicated with, e.g., a deep learning transformer. On the cognitive side, the deep learning model would need to implement an equivalent of chunking (including its timing aspects). On the evolutionary side, the deep learning model may need training data in the form of cognitive models (which may be difficult as these are not known in advance), or utilise reinforcement learning. Alternatively, a deep learning model may be used to replace only one of GEVL's components (i.e., either cognitive *or* evolutionary). The advantage of genetic programming is that it is efficient (for the types of problems discussed in the current paper) and that it works directly from the grammatical definitions of the models. However, it is an open question of how it would compare to other approaches – for the current, as well as the more complex tasks.

One potential criticism is that our models produced suboptimal learning strategies in order to fit the longer durations of human learning. There are two ways of answering this criticism. On the one hand, this may indeed be unrealistic and deserves further investigation. On the other hand, suboptimal learning routines have long been known in psychology – e.g., as "satisficing" (Simon, 1991). Moreover, the suboptimal attention shifts displayed by GEVL may be in line with research into saccadic eye movement and the underlying attention function (Cajar et al., 2016).

Another potential criticism of the current modelling approach is "overfitting" – changing free parameters to achieve better fit may lead to poor generalisability beyond the currently simulated data (Tetko, Livingstone, & Luik, 1995).

This study followed the advice of Simon (1992) and attempted to address the issue by doubling the *data explained/free parameters used* ratio – the same free parameters were used for both pattern similarity and constant learning tasks. One future extension to the current study would be to replicate EPAM's simulation of other verbal learning experiments without resorting to hand-crafting task-specific strategies. Potential use of other cognitive architectures (e.g., ACT-R, SOAR, LIDA and so on) in combination with the GEMS environment is similarly intriguing, as is "cross-breeding" of operators that may come from multiple architectures to simultaneously populate a single GEMS evolutionary pool.

Finally, it is important to emphasise that the seemingly trivial and contrived verbal learning simulations are not an end in themselves. As was discussed above, the basic mechanisms that were established in verbal learning and other similar experimental paradigms helped to develop various cognitive models that subsequently went on to simulate highly complex human behavior in multiple domains (e.g., Bennett et al., 2020; Freudenthal et al., 2016). We anticipate that this trend will continue with GEVL.

To conclude, our study further integrates genetic/evolutionary aspects with cognitive models (thus bridging the "how?" and "why?" questions) and automates task-specific strategy discovery. Our findings offer further support to the mechanisms proposed by chunking theory, connect them to the evolutionary approach, and make further inroads towards a Unified Theory of Cognition (Newell, 1990).

References

Bartlett, L., Pirrone, A., Javed, N., Lane, P., & Gobet, F. (2023). Genetic programming for developing simple cognitive models. In F. K. A. M. Goldwater, B. K. Hayes, & D. C. Ong (Ed.), *Proceedings of the 45th annual conference of the cognitive science society*. Sydney, Australia.

Bennett, D., Gobet, F., & Lane, P. (2020). Forming concepts of Mozart and Homer using short-term and long-term memory: A computational model based on chunking. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 178-184). Toronto.

Brave, S. (1996). *The evolution of memory and mental models using genetic programming*. Paper presented at the Proceedings of the 1st annual conference on genetic programming, Stanford, California.

Bugelski, B. R. (1962). Presentation time, total time, and mediation in paired-associate learning. *Journal of Experimental Psychology, 63*, 409-489.

Cajar, A., Schneeweiß, P., Engbert, R., & Laubrock, J. (2016). Coupling of attention and saccades when viewing scenes with central and peripheral degradation. *Journal of Vision, 16*(2), 8-8. doi:10.1167/16.2.8

Chase, W. G., & Simon, H. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Crespi, B. (2014). The insectan apes. *Human Nature, 25*(1), 6-27.

Crespi, B. (2017). Shared sociogenetic basis of honey bee behavior and human risk for autism. *Proceedings of the National Academy of Sciences, 114*(36), 9502-9504. doi:10.1073/pnas.1712292114

Dawkins, R. (1976). *The selfish gene.* Oxford: Oxford University Press.

Dawkins, R. (1982). *The extended phenotype: The long reach of the gene.* Oxford: Oxford University Press.

Feigenbaum, E. A. (1959). *An information processing theory of verbal learning.* (P-1817). Santa Monica, CA

Feigenbaum, E. A. (1963). The simulation of verbal learning behavour. *Proceedings of the Western joint computer conference, 19*, 121-132.

Feigenbaum, E. A., & Simon, H. (1962). A theory of the serial position effect. *British Journal of Psychology, 53*(3), 307.

Feigenbaum, E. A., & Simon, H. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8*, 305-336.

Freudenthal, D., Pine, J., Jones, G., & Gobet, F. (2016). Developmentally plausible learning of word categories from distributional statistics. In D. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *38th annual conference of the cognitive science society*. Austin, TX.

Gobet, F. (1993). A computer model of chess memory. In W. Kintsch (Ed.), *Fifteenth annual meeting of the cognitive science society* (pp. 463-468): Erlbaum.

Gobet, F. (2000). *Discrimination nets, production systems and semantic networks: Elements of a unified framework*. Evanston: The Association for the Advancement of Computing in Education.

Gobet, F., & Lane, P. (2012). Chunking mechanisms and learning. In M. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 541-544). New York: NY: Springer.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*(6), 236.

Gobet, F., Lloyd-Kelly, M., & Lane, P. (2016). What's in a name? The multiple meanings of "chunk" and "chunking". . *Frontiers in Psychology, 7*.

Gobet, F., & Simon, H. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science, 24*, 651-682.

Gregg, L. W., & Simon, H. (1967). An information processing explanation of one-trial and incremental learning. *Journal of verbal learning and verbal behavior, 6*, 780-787.

Gunaratne, C., & Patton, R. (2022). *Genetic programming for understanding cognitive biases that generate polarization in social networks*. Paper presented at the Proceedings of the Genetic and Evolutionary Computation Conference

Companion, Boston, MA. https://doi.org/10.1145/3520304.3529069

Hintzman, D. L. (1968). Explorations with a discrimination net model for paired associate learning. *Journal of Mathematical Psychology, 5*, 123-126.

Hintzman, D. L. (1969). Backward recall as a function of stimulus similarity. *Journal of verbal learning and verbal behavior, 8*, 384-387.

Hunter, P. (2018). The revival of the extended phenotype. *EMBO reports, 19*(7), e46477. doi:10.15252/embr.201846477

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *The Artificial Intelligence review, 53*(1), 17-94. doi:10.1007/s10462-018-9646-y

Lane, P., Cheng, P. C.-H., & Gobet, F. (2000). CHREST + : Investigating how humans learn to solve problems using diagrams. *AISB Quarterly, 103*, 24–30.

Lane, P., & Gobet, F. (2012). Using chunks to categorise chess positions. In M. Bramer & M. Petrides (Eds.), *Specialist group on artificial intelligence international conference 2012* (pp. 93-106). London: Springer-Verlag.

Lane, P., & Gobet, F. (2013). Evolving non-dominated parameter sets for computational models from multiple experiments. *Journal of Artificial General Intelligence, 4*, 1-30. doi:10.2478/jagi-2013-0001

Lane, P., Sozou, P. D., Addis, M., & Gobet, F. (2014). Evolving process-based models from psychological data using genetic programming.

Lieto, A. (2019). Heterogeneous proxytypes extended: Integrating theory-like representations and mechanisms with prototypes and exemplars. In A. Samsonovich (Ed.), *Biologically inspired cognitive architectures 2018* (Vol. 848, pp. 217-227). London: Springer.

Mathy, F., Fartoukh, M., Gauvrit, N., & Guida, A. (2016). Developmental abilities to form chunks in immediate memory and its non-relationship to span development. *Frontiers in Psychology, 7*, 201.

Mccrary, J. W., & Hunter, W. S. (1953). Serial position curves in verbal learning. *Science, 117(3032),* 131.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA, US: Harvard University Press.

Richman, H. B., Gobet, F., Staszewski, J., & Simon, H. (1996). Perceptual and memory processes in the acquisition of expert performance: The EPAM model. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games* (pp. 167–187). Mahwah, MA: Erlbaum.

Richman, H. B., & Simon, H. (1989). Context effects in letter perception: Comparison of two theories. *Psychological Review, 96*(3), 417-432. doi:10.1037/0033-295X.96.3.417

Richman, H. B., Simon, H., & Feigenbaum, E. A. (2002). Simulations of paired associate learning using EPAM VI. *Unpublished*.

Richman, H. B., Staszewski, J. J., & Simon, H. (1995). Simulation of expert memory using EPAM IV. *Psychological Review, 102*(2), 305-330.

Ridley, M. (2016). In retrospect: The selfish gene. *Nature, 529*(7587), 462-463. doi:10.1038/529462a

Rock, I. (1957). The role of repetition in associative learning. *Amercan journal of psychology, 70*, 186-193.

Rubenstein, D. R., & Hofmann, H. A. (2015). Proximate pathways underlying social behavior. *Current Opinion in Behavioral Sciences, 6*, 154-159.

Schiller, M., & Gobet, F. (2014). Cognitive models of gambling and problem gambling. In F. Gobet & M. R. G. Schiller (Eds.), *Problem gambling: Cognition, prevention and treatment* (pp. 74-103). London: Palgrave Macmillan.

Simon, H. (1974). How big is a chunk? *Science, 183*(4124), 482-488.

Simon, H. (1991). *The sciences of the artificial.* New York: MIT Press.

Simon, H. (1992). What is an "explanation" of behavior? *Psychological Science, 3*, 150-161.

Simon, H., & Gilmartin, K. (1973). A simulation of memory for chess positions. . *Cognitive Psychology, 5*, 29-46.

Smith, R., Gobet, F., & Lane, P. (2007). *An investigation into the effect of ageing on expert memory with CHREST.* Paper presented at the Proceedings of The Seventh UK Workshop on Computational Intelligence, Aberdeen.

Tetko, I. V., Livingstone, D. J., & Luik, A. I. (1995). Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences, 35*(5), 826.

Thune, L. E., & Underwood, B. J. (1943). Retroactive inhibition as a function of degree of interpolated learning. *Journal of Experimental Psychology, 32*, 185-200.

Underwood, B. J. (1953). Studies of distributed practice: Viii. Learning and retention of paired nonsense syllables as a function of intralist similarity. *Journal of Experimental Psychology, 45*.

From States to Transitions: Discrete Time Markov Chains for Affect Dynamics

Francesca Borghesi (francesca.borghesi@unito.it)
Pietro Cipresso (pietro.cipresso@unito.it)
Department of Psychology, University of Turin, 10124, Italy

Abstract

Affect dynamics, or the study of changing patterns of emotional responses across time, has emerged as a key field of research in Mathematical Psychology. Traditionally, Affect dynamics research has relied on the Experience Sampling Method (ESM), a data gathering technique in which participants describe their feelings, thoughts, and behaviors at various times throughout the day. This technique studies Intensive Longitudinal Data (ILD) using Mixed Linear or Nonlinear Models (MLM) or Vector Autoregressive Models (VARs) (VAR). These theories characterize emotion in terms of time and complexity. However, they fail to recognize the underlying unity of emotional dynamism: the transition between affects. Although emotions follow one another, the transition only considers the dyadic relationship between the current state and the immediately following future state. In this paper, we will show how to use and implement Discrete Time Markov Chains to evaluate each transition between the current and the future emotional states, while neglecting earlier transitions. Researchers can use Markov chains to quantify the likelihood of transitions from one emotional state to another over time, allowing a short-term understanding of the dynamics of affect.

Keywords: Affect dynamics; Markov Chain; stochastic model; Russell circumplex model; psychometrics

Introduction

Affect dynamics, or the study of the changing patterns of emotional experiences throughout time, has emerged as a critical topic of inquiry in the psychological sciences (Hamaker et al., 2015; Waugh & Kuppens, 2021). This discipline studies how emotions evolve and interact to influence human behavior, decision-making, and well-being (Puccetti et al., 2021). Traditional models have explored affect dynamics from a variety of perspectives, including psychological theories of emotion regulation, computing models mimicking emotional states, and statistical methods for examining temporal patterns of affective experiences (Borghesi, Chirico, et al., 2023; Borghesi, Murtas, Mancuso, et al., 2023; Lazarus et al., 2021).

Historically, the field relied on linear and time-invariant models to convey the essence of emotional shifts. Despite the progress made with these current models, one fundamental difficulty remains: accurately capturing the complexity and non-linear nature of emotional dynamics. Emotions, by definition, are dynamic and fleeting, making them difficult to categorize or progress linearly. Traditionally, affect dynamics analysis has relied on the Experience Sampling Method (ESM), a data gathering technique in which participants describe their feelings, thoughts, and behaviors at random times throughout the day. This method has produced useful longitudinal insights into self-reported emotions at precise time intervals, revealing emotional variability in response to various contexts and stimuli. The mathematical modeling most used are Mixed Linear Model (MLM) or Autoregressive Models (AR) and Vector Autoregressive Models (VAR). These models consider the long-term history of emotions, describing in a way more the concept of mood dynamics.

Hence, to investigate the short-term affect transition, we propose an implementation of Discrete Time Markov Chains. Markov chains allow for the modeling of emotional state transitions as stochastic processes, with the likelihood of shifting from one state to another set irrespective of observable or unobserved transitions. This strategy considers both the emotional changes that participants reported and the fact that some changes did not occur at all. Thus, Markov chains open up new avenues for understanding emotional complexity, bypassing the constraints of standard experimental models and providing a strong instrument for investigating the true structure of human emotional experiences. Enter Markov chains, a mathematical construct that exists at the intersection of probability theory and stochastic processes. Markov chains offer a solid framework for simulating random processes in which the future state depends solely on the present state and not on the sequence of events that came before it. This trait, known as the Markov property, makes Markov chains ideal for modeling the sequential and stochastic nature of emotional transitions. The application of Markov chains to affect dynamics is a novel step forward, providing a new lens through which to investigate the complex web of emotional states. Markov chains' capacity to simulate complicated stochastic processes enables the analysis of emotional transitions in a granular and flexible manner in response to the changing nature of affective experiences.

The use of Markov chains in the study of affect dynamics emphasizes not only the interdisciplinary nature of modern psychology research but also the potential for mathematical models to provide new insights into the human emotional experience. Using Markov chain principles, researchers may

build sophisticated models that anticipate emotional states across time, providing a dynamic view of emotional evolution that reflects the complexity and variety inherent in human affect.

Statistical Models in Affect Dynamics: A Journey from Intensive Longitudinal Model to Markov Chains

Intensive Longitudinal Data (ILD) is used to study affective processes to capture significant dynamics. Depending on the process, daily, moment-to-moment, or even second-to-second measurements may be required (Bolger, Davis, & Rafaeli, 2003; Trull & Ebner-Priemer, 2013). These measurements can be obtained through a daily diary, ambulatory assessment, experience sampling, observations, or laboratory measurements. Smartphones, accelerometers, and intelligent shirts have simplified the collection of ILD, making intensive longitudinal studies a viable alternative to traditional research methods like cross-sectional and panel.

The ILD range considers the data's scope (individual vs. group) and complexity (single vs. multiple variables), to the nature of the affective processes (stable vs. changing, linear vs. complex interactions), the temporal aspect of data collection (discrete vs. continuous timing), the type of variables involved (categorical vs. numerical), the analytical approach (time-based vs. frequency-based analysis), and the focus of the analysis (process modeling vs. summary statistics).

Hence, the mathematical modeling most used are Mixed Linear Model (MLM) or Autoregressive Models (AR) and Vector Autoregressive Models (VAR).

MLM also known as hierarchical linear models, allow for modeling data that come from multiple levels of grouping (for example, repeated measurements, which are the lower levels, nested within individuals, which are the higher levels). In an MLM, changes over time can be modelled as fixed effects, while variability between individuals can be captured as random effects.

The basic formula for an MLM for ILD might be:

$$y_{it} = \beta_0 + \beta_1 \times \text{Time}_{it} + u_{0i} + u_{1i} \times \text{Time}_{it} + \epsilon_{it}$$

where y_{it} is the outcome for individual i at time t, β_0 and β_1 are the coefficients of fixed effects, u_{0i} and u_{1i} are the random effects for individual i, and ϵ_{it} is the residual error.

These models are used to analyze time series and are particularly useful for modeling the temporal dependence between observations. In an AR model, the current measurement is predicted from past measurements. In a VAR model, this approach is extended to multiple temporal variables, allowing for examining how each variable is influenced by its own past measurements as well as by the

past measurements of other variables. An AR model of order p, AR(p), for a time variable y at time t is defined as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

where:
 c is a constant,
 $\phi 1, \phi 2, ..., \phi_p$ are coefficients representing the influence of past observations on the current observation,
 ε_t is the error term (white noise) at time t.

These models capture the temporal dependency within a single time series, allowing for analyzing how past values influence future values.

Vector autoregressive (VAR) Models extend the autoregressive approach to multivariate systems. VAR model of order p for a vector of k time variables $\mathbf{y}_t$ is defined as:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + \varepsilon_t$$

where:
 c is a vector of constants, $A_1, A_2, ..., A_p$ are matrices of coefficients representing the influence of past observations on the current vector of observations, ε_t is a vector of multivariate error terms at time t.

When considering multivariate linear models, which consider the ratings of multiple emotions to explain and impact previous ones, a complex relational model is created. This model might obscure the concept of transitions between emotional states since it encompasses a network of interactions rather than clear-cut transitions. The proposal to analyze emotional transitions specifically, considering pairs of emotions and employing Discrete Time Markov processes, aims to address this gap. This approach allows to capture the probabilistic nature of these transitions based on the current state without the necessity for the preceding states to directly influence the future beyond the immediate past.

The Discrete Time Markov processes model, distinct from AR and VAR models or MLM, provides a framework suitable for examining the series of emotional transitions, as it relies on the assumption that the future emotional state depends only on the present state. This aligns with the intuitive understanding of emotional transitions where each state is a discrete event influenced by the immediate past, enabling a clearer analysis of the emotional trajectory of an individual. By focusing on pairs of emotions (or duplexes in mathematical terms) and applying discrete Markovian processes, the analysis can capture the essence of emotional transitions. This approach respects the temporal sequence of emotions while providing specific insights into the dynamics of emotional transitions, thus offering a more tailored understanding of affect dynamics in relation to mental flexibility and environmental stimuli.

Mathematical Model of Discrete Time Markov Chains

A Discrete Time Markov Chain is a mathematical model used to describe a process where a system transitions from one state to another in a sequence of steps that occur at discrete

points in time. This model is characterized by a series of random variables X_1, X_2, X_3, …, each representing the state of the system at a specific time step. The key feature of a Discrete Time Markov Chain is the Markov property, which states that the probability of transitioning to any future state depends only on the current state and not on the sequence of events that preceded it. Mathematically, this property can be expressed as:

$$P(X_{n+1}=x \mid X_1=x_1, X_2=x_2, … , X_n=x_n) = P(X_{n+1}=x \mid X_n=x_n)$$

for any state x and any sequence of states $x_1,x_2,…,x_n$, where n represents the step number in the Markov chain. This means that the future behavior of the process can be determined entirely by its present state, making the system's history irrelevant for predicting its future state transitions. The transition probabilities between states are typically represented in a matrix form, known as the transition matrix, where each entry indicates the probability of moving from one state to another.

In the exploration of stochastic processes, discrete Markov chains stand out for their unique property of time homogeneity and memory lessness. This fundamental characteristic assert that the probability of transitioning from one state to another solely depends on the current state and not on the sequence of events that preceded it.

A Discrete Time Markov Chain is defined by a set of states, S, and a transition probability matrix, P. Time homogeneity is encapsulated in the axiom:

$$P(X_{n+1}=j \mid X_n=i) = P(X_1=j \mid X_0=i)=P_{ij}$$

for all states i, j in S, and for all n in natural numbers. This implies that the probability of transitioning from state i to state j remains constant over time. The transition matrix P, where each element P_{ij} represents the probability of moving from state i to state j, is central to the description and analysis of a Markov chain.

The initial state of the Markov chain is described by a probability distribution vector π_0, where $\pi_{0(i)} = P(X_0 = i)$. The vector π_0 encapsulates the probabilities associated with each state at the beginning of the process, a priori probability.

The evolution of a Markov chain is characterized by the progression of states over discrete time steps, called steady state (π_n). The transition probabilities dictate the dynamics of this stochastic process.

Mathematically, the state distribution at step n, denoted as π_n, is obtained through the relation:

$$\pi_n=\pi_0 P^n$$

where P^n represents the nth power of the transition matrix P. This fundamental relationship underscores the process's discrete nature and the Markov property. It highlights how the initial distribution, combined with the transition matrix, dictates the state probabilities at any future step.

A cornerstone theorem in the theory of Markov chains states that any irreducible and aperiodic Markov chain, defined on a finite state space S with a stochastic transition matrix P, converges to a unique stationary distribution π, where:

$$\lim_{n \to \infty} \left(P_{ij}\right)^n = \pi_j \; \forall \; i,j \in S$$

This theorem underscores the long-term behaviour of Markov chains, revealing that, regardless of the initial state, the chain converges to a stationary distribution, illustrating the system's equilibrium properties.

Furthermore, the characteristics of Markov chains such as irreducibility, periodicity, and recurrence are critical in determining the long-term behavior. A Markov chain is irreducible if every state is reachable from every other state, aperiodic if the greatest common divisor of the lengths of all possible loops for each state is one, and a state is recurrent if it is guaranteed to return to itself within a finite number of steps with probability one.

Application of Markov Chain in Affect Dynamics

The application of Markov chains in the field of affect dynamics offers a novel approach to understanding the complex nature of emotional state transitions over time. By representing affective states as discrete, distinguishable states within a Markov model, researchers can analyze the probabilities associated with transitioning from one emotional state to another. This approach not only captures the stochastic nature of emotional changes but also aligns with the temporal granularity of affective fluctuations observed in empirical data (Borghesi, Chirico, et al., 2023; Cipresso et al., 2023).

Considering the Russell Circumplex model, affect can be categorized into four quadrants based on valence and arousal dimensions. This approach breaks down emotions into finer nuances, situating them within a two-dimensional space of valence (pleasant-unpleasant) and arousal (activated-deactivated). Each quadrant represents a unique blend of these two dimensions: high arousal-positive valence (happy, excited), high arousal-negative valence (angry, anxious), low arousal-positive valence (relaxed, content), and low arousal-negative valence (sad, depressed) (Posner et al., 2005; Russell, 2017). For example, if a person is currently in a state of high arousal and positive valence, the Markov chain can predict the likelihood of their next state based on historical transition data. Will they remain in this happy state, or transition to a more relaxed state (low arousal-positive valence), or perhaps shift to a high arousal-negative valence state due to an external stressor? By quantifying the transition probabilities between these states, Discrete Markov chains offer a mathematical model to capture the fluidity and dynamism of human emotions. Considering the four distinct affective states, namely stress (A), engagement (B), boredom (C), and relax (D) reflecting a separate emotional condition,

emerging 12 possible transitions, divided in horizontal, vertical and oblique one, as Figure 1 demonstrates.
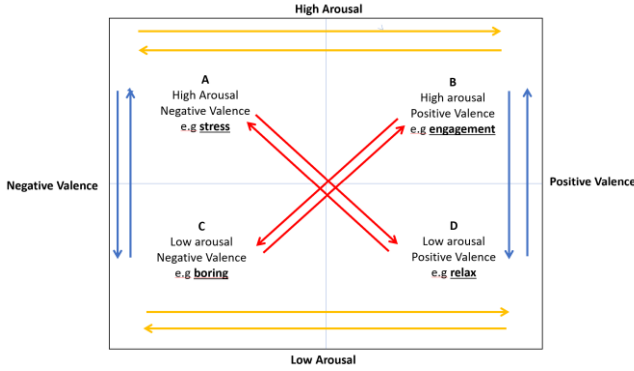

Figure 1: Transitions of 4 affect states: AB, BA, CD, DC, AC, CA, BD, DB, AD, DA, BC, CB

In this case, the Markov chain is represented by a transition matrix P, which is a 4x4 matrix where each element p_{ij} represents the probability of transitioning from state i to state j. For our four emotional states, the transition matrix is:

$$P = \begin{bmatrix} p_{AA} & p_{AB} & p_{AC} & p_{AC} \\ p_{BA} & p_{BB} & p_{BC} & p_{BD} \\ p_{CA} & p_{CB} & p_{CC} & p_{CD} \\ p_{DA} & p_{DB} & p_{DC} & p_{DD} \end{bmatrix}$$

Here, each p_{ij} is a probability, so each element in the matrix must satisfy $0 \leq p_{ij} \leq 1$, and the sum of the probabilities in each row must equal 1, reflecting the total probability of transitioning from any given state to all possible states, including remaining in the same state.

Markovization Process

Incorporating questionnaire responses or physiological data into a Markov chain involves a methodical approach to quantify affect transitions. This process involves converting raw data into a format compatible with the stochastic nature of Markov chains, which we achieve using normalization indexes. Following, a formal and mathematical example to elucidate how this could be implemented.

Consider conducting an Experience Sampling Method (ESM) study over seven days, querying participants five times daily about their current emotional state. Given the four defined states (A, B, C, D), participant can respond with one of these states at each query, resulting in a total of 43 responses. First, compile the responses into a frequency matrix, counting the number of transitions from each state to every other state, including self-transitions. The frequency matrix F might look like this:

$$F = \begin{bmatrix} f_{AA} & f_{AB} & f_{AC} & f_{AC} \\ f_{BA} & f_{BB} & f_{BC} & f_{BD} \\ f_{CA} & f_{CB} & f_{CC} & f_{CD} \\ f_{DA} & f_{DB} & f_{DC} & f_{DD} \end{bmatrix}$$

where f_{ij} represents the frequency of transitions from state i to state j observed in the study. Here a subject example matrix, in which we count the number of transitions (Table 1):

Table 1: Absolute transition matrix

	Engagement	Stress	Relax	Boring	Total row
Engagement	16	1	0	0	17
Stress	1	0	1	1	3
Relax	3	2	16	1	22
Boring	0	0	1	0	1

Next, normalize this matrix to convert frequencies into probabilities. This is done by dividing each element in a row by the total sum of elements in that row, resulting in the transition matrix P:

$$F = \begin{bmatrix} \frac{f_{AA}}{\Sigma_j f_{Aj}} & \frac{f_{AB}}{\Sigma_j f_{Aj}} & \frac{f_{AC}}{\Sigma_j f_{Aj}} & \frac{f_{AD}}{\Sigma_j f_{Aj}} \\ \frac{f_{BA}}{\Sigma_j f_{Bj}} & \frac{f_{BB}}{\Sigma_j f_{Bj}} & \frac{f_{BC}}{\Sigma_j f_{Bj}} & \frac{f_{BD}}{\Sigma_j f_{Bj}} \\ \frac{f_{CA}}{\Sigma_j f_{Cj}} & \frac{f_{CB}}{\Sigma_j f_{Cj}} & \frac{f_{CC}}{\Sigma_j f_{Cj}} & \frac{f_{CD}}{\Sigma_j f_{Cj}} \\ \frac{f_{DA}}{\Sigma_j f_{Dj}} & \frac{f_{DB}}{\Sigma_j f_{Dj}} & \frac{f_{DC}}{\Sigma_j f_{Dj}} & \frac{f_{DD}}{\Sigma_j f_{Dj}} \end{bmatrix}$$

This resulting matrix P is the stochastic matrix for the Markov chain, with each element p_{ij} representing the probability of transitioning from state i to state j. To smooth data, reducing the impact of minor variations and eliminating zeros that might indicate improbable transitions or missing data, a Dirichlet smoothing was applied in the normalization process. α_j is the smoothing parameter for state j: $\alpha = 0.5$ for more frequent states (e.g. the total row of Engagement and Relax) $\alpha = 0.1$ for less frequent states (e.g. the total row of Stress and Boring). Hence, the normalization takes account of the α parameters and absolute frequency of transition:

$$P_{ij} = \frac{f_{ij} + \alpha_j}{N_i + \sum_j \alpha_j}$$

Here the final probability transition matrix (Table 2):

Table 2: Probability transition of matrix

	Engagement	Stress	Relax	Boring	Total row
Engagement	0.868	0.079	0.026	0.026	1.00
Stress	0.324	0.029	0.324	0.324	1.00
Relax	0.146	0.104	0.688	0.063	1.00
Boring	0.071	0.071	0.786	0.071	1.00

In addressing physiological data for constructing Markov chains, we encounter a substantial challenge: transitioning

from the categorical data of questionnaires to the continuous measurements typical of physiological parameters. In this context, the laboratory approach becomes essential as it allows for precise measurement of induced affective state transitions.

Consider, for example, the use of image blocks from the International Affective Picture System (IAPS) or an equivalent emotional elicitation tool. In an experimental design, we can induce measurable affective transitions through changes in image blocks. Unlike the Experience Sampling Method (ESM), where the state transition may remain uncertain, in a controlled environment, we can accurately identify when and how each affective transition occurs. To translate these data into a Markov matrix, we proceed as follows:

- Temporal Segmentation: Define specific periods for each image block and for transitions between blocks. For instance, we might consider the last few seconds of a block as the transition moment and the middle moments of the blocks as representative of stable affective states.
- Measurement and Calculation of Indexes: For each temporal segment, we calculate relevant physiological indices such Standard Deviation (SD), Relative Standard Deviation (RSD), Root Mean Square of Difference (RMSSD) proposed by Pirla et al., (2023). These indices should be calculated separately for each transition period and each stable affective state.
- Normalization and Markovization: We convert the index values into relative probabilities to reflect the likelihood of transitioning from one affective state to another. For example, if an index shows a significant change between the end of one affective state and the beginning of another, this indicates a higher transition probability, which can be normalized across all observed transitions.
- Construction of the Transition Matrix: We organize the normalized probabilities into a Markov transition matrix, where each row represents an initial affective state and each column an arriving affective state. The sum of probabilities in each row will equal 1, reflecting the stochastic nature of the transitions.

This approach provides a newly mathematically framework for affective transitions, allowing for precise analysis of emotional dynamics in response to controlled stimuli.

Affect Markov Indexes

From the Markovian chain model, which encapsulates the dynamics of affect transitions as informed by both subjective experiences and physiological responses, a multitude of indexes can be extracted to analyze and understand the nuances of these transitions. These indexes, derived from the Markov transition matrix, provide both descriptive and predictive insights into affect dynamics. The descriptive one includes probabilities for vertical, horizontal, and oblique transitions, each representing different types of emotional changes. These transitions are categorized as follows:

- Vertical transitions (involving changes primarily in the arousal dimension without significant changes in valence): $p_{AC}, p_{CA}, p_{BD}, p_{DB}$ (e.g., AC from stress to boredom or BD from engagement to relax).
- Horizontal transitions (involving changes primarily in the valence dimension without significant changes in arousal): $p_{AB}, p_{BA}, p_{CD}, p_{DC}$ (e.g., AB from stress to engagement or CD from boredom to relax).
- Oblique transitions (involving simultaneous changes in both arousal and valence dimensions): $p_{AD}, p_{DA}, p_{BC}, p_{CB}$ (e.g., AD from stress to relax or BC from engagement to boredom).

Additionally, in the context of Markov chains, we had also to consider internal variability within each affective state, corresponding to the transitions within the same state (e.g., AA, BB, CC, DD), also named *state trait transition*.

However, the Markov matrix not only provides descriptive indices of the transitions, but it can also provide predictive indices, which explain after n steps (hypothetically n state transitions) with what probability the subject will be in one of the different states, also called *steady states* (π_n). The steady state distribution provides valuable insights into the long-term behavior of the system, offering insights into the equilibrium distribution of states after a large number of transitions.

The steady states depend by two factors: the initial state vector (π_0) and the subject-specific transition matrix. The initial state vector (π_0) posits an a priori probability of the subject's presence in one of the four quadrants, serving as the starting point for the Markov chain: i.e. the subject before being subjected to stimuli is equally likely to be in one of the four affective states. Our initial state vector considered equiprobability between initial affective states. Through iterative multiplication (e.g., 10 steps) of the initial states vector and the transition matrix, we arrive at the steady states. Thus, the steady states serve as an updated version of the original matrix, modifying the initial equilibrium based on the empirical affective transitions experienced by the subject: They describe the likelihood of discovering one of the four states after attempting ten different hypothetical transitions. The iterative process between the initial states vector of probability and the transition matrix, culminating in the steady states, underscores the dynamic interplay between predisposition and experience in shaping the affective journey of individuals.

Conclusion

Markov processes provide the concentrated and exact modeling of transitions between emotional states, making them ideal for understanding how people go from one emotion to another with a focus on the current transition, regardless of previous sequences. Autoregressive (AR) and Vector Autoregressive (VAR) models can capture temporal dynamics, however they may not accurately characterize precise transitions between emotional states. In contrast, Markov processes simplify the analysis by confining the reliance to the immediately previous states, making the model more understandable and interpretable in the emotional context. The finding that people can experience numerous emotions at the same time, with one predominating and affecting the next, is consistent with the Markovian process theory. These models reflect the probability-based character of emotional transitions, stressing the present emotion's importance in deciding the future state. Using Markov processes to study pairs of emotions allows for a more individualized and thorough investigation of people's emotional trajectories, which may be very useful in clinical and psychological research settings.

However, it is vital to remember that every model has limits, and a model's applicability is determined by the study's unique aims and the nature of the data. Furthermore, the use and interpretation of Markov processes require a precise characterization of emotional states as well as reliable data gathering. Overall, the suggested strategy is well-founded and promises to deliver unique insights into emotional dynamics, signifying a substantial step forward over more standard methodologies in future research initiatives. Future studies could significantly benefit from implementing Hidden Markov Models (HMMs) to delve deeper into the complexities of emotional transitions. Unlike traditional Markov models that directly correlate observed emotional states to transition probabilities, HMMs introduce an additional layer of abstraction by positing that observable behaviors are influenced by hidden, unobserved states. Finally, the use of continuous stimuli, such as those provided by Virtual Reality, offers a promising avenue for studying affect dynamics under controlled, yet highly immersive and realistic conditions. VR technology enables the simulation of complex, dynamic environments where users can experience a wide range of emotionally evocative scenarios in a safe and controlled setting. This immersive approach allows for the continuous monitoring of emotional responses to nuanced and evolving stimuli, which can be particularly beneficial for understanding how individuals navigate through emotionally charged environments or situations (Borghesi, Murtas, Mancuso, et al., 2023; Borghesi, Murtas, Pizzolante, et al., 2023).

References

Borghesi, F., Chirico, A., & Cipresso, P. (2023). Outlining a novel psychometric model of mental flexibility and affect dynamics. *Frontiers in Psychology*, *14*, 1183316. https://doi.org/10.3389/FPSYG.2023.1183316/BIBTEX

Borghesi, F., Murtas, V., Mancuso, V., & Chirico, A. (2023). *Continuous Time Elicitation Through Virtual Reality to Model Affect Dynamics*. 258–276. https://doi.org/10.1007/978-3-031-49368-3_16

Borghesi, F., Murtas, V., Pizzolante, M., Chirico, A., & Cipresso, P. (2023). Affect Dynamics through Virtual Reality. *ANNUAL REVIEW OF CYBERTHERAPY AND TELEMEDICINE* , 11–13.

Cipresso, P., Borghesi, F., & Chirico, A. (2023). Affects affect affects: A Markov Chain. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1162655

Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., & Tuerlinckx, F. (2015). Modeling Affect Dynamics: State of the Art and Future Challenges. *7*(4), 316–322. https://doi.org/10.1177/1754073915590619

Lazarus, G., Song, J., Crawford, C. M., & Fisher, A. J. (2021). A Close Look at the Role of Time in Affect Dynamics Research. *Affect Dynamics*, 95–116. https://doi.org/10.1007/978-3-030-82965-0_5

Pirla, S., Taquet, M., & Quoidbach, J. (2023). Measuring affect dynamics: An empirical framework. *Behavior Research Methods*, *55*(1), 285–300. https://doi.org/10.3758/S13428-022-01829-0/FIGURES/7

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734. https://doi.org/10.1017/S0954579405050340

Puccetti, N. A., Villano, W. J., & Heller, A. S. (2021). The Neuroscience of Affective Dynamics. *Affect Dynamics*, 33–60. https://doi.org/10.1007/978-3-030-82965-0_3

Russell, J. A. (2017). Mixed Emotions Viewed from the Psychological Constructionist Perspective: *Http://Dx.Doi.Org/10.1177/1754073916639658*, *9*(2), 111–117. https://doi.org/10.1177/1754073916639658

Waugh, C. E., & Kuppens, Peter. (2021). *Affect Dynamics* (C. E. Waugh & P. Kuppens, Eds.). Springer International Publishing. https://doi.org/10.1007/978-3-030-82965-

Intractability Obstacles to Explanations of Communication

Laura van de Braak (laura.vandebraak@donders.ru.nl)
Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Ronald de Haan (me@ronalddehaan.eu)
Institute for Logic, Language and Computation (ILLC), University of Amsterdam, The Netherlands

Mark Dingemanse (mark.dingemanse@ru.nl)
Center for Language Studies, Radboud University, The Netherlands

Ivan Toni (ivan.toni@donders.ru.nl)
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Iris van Rooij (iris.vanrooij@donders.ru.nl)
Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands
Department of Linguistics, Cognitive Science, and Semiotics & Interacting Minds Centre, Aarhus University, Denmark

Mark Blokpoel (mark.blokpoel@donders.ru.nl)
Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Abstract

Even when talking about novel things and without a fully shared vocabulary, people can come to understand each other through communicative turn taking (what we call *communicative alignment*). State-of-the-art computational models cannot yet explain this capacity, because (1) empirically corroborated models only work under shared knowledge and vocabularies, and leave out interactive processes needed to overcome misalignment; (2) models that do include misalignment and interactive processes cannot account for communicative successes under real-world conditions; and (3) models that overcome the limits in (2) use a theoretical 'hack'. In this paper, we add a challenge to the list: the interactive processes in both models of type (2) and (3) are intractable. We explore the robustness and implications of this theoretical challenge for models of communicative alignment in general.

Keywords: communication; interaction; computational complexity; intractability

Introduction

Consider the fruit platter in Figure 1. Now imagine a friend asks you: "Could you pass me the dragon fruit?" If you have never seen a dragon fruit before you may have no idea what they mean. In an attempt to understand them, you could inquire, "Is that the prickly fruit in the bowl?" In that case, they may respond: "No, I mean the pink one." After this exchange, you know which fruit they want and in future contexts you will likely also understand their request.

People's capacity to interactively come to mutual understanding (Dingemanse et al., 2015; Schegloff, Jefferson, & Sacks, 1977; H. H. Clark & Schaefer, 1987), even in the absence of a fully shared vocabulary (Quine, 2013), is remarkable.[1] We call this capacity *communicative alignment*. It has

Figure 1: A fruit platter with potentially unfamiliar fruits.

been experimentally demonstrated that communicative alignment is possible even when people talk about completely unfamiliar objects (e.g., using the Tangram task (H. H. Clark & Wilkes-Gibbs, 1986; Holler & Wilkin, 2011) and the Fribbles (Barry, Griffith, De Rossi, & Hermans, 2014; Eijk et al., 2022); see Box 1). So far, this feat defies a computational explanation for several reasons.

First, empirically corroborated computational models of pragmatic communication only model a slice of the capacity, i.e., only one utterance ("Could you pass me the dragon fruit") and one inference ("Is that the prickly one?")[2]

---

[1]We conceive of this capacity as something that people can do

under minimal enabling conditions. For instance, these conditions include that both speakers commit to trying to reach mutual understanding and are engaging in good faith. If one of the actors violates these conditions (e.g., by engaging in ill-faith, deception (Dynel, 2020), or epistemic injustice (Pohlhaus, 2012)), then successful communicative alignment may be blocked. This does not mean that the capacity for communicative alignment is not real or robust, but it does need minimal conditions to operate properly.

[2]In fact, these models will not even capture this slice intact. The

(Hawkins, Frank, & Goodman, 2017), whereas communicative alignment may require a series of interactions before mutual understanding is achieved. Second, recent extensions of such models that attempt to include communicative turn-taking do not yet show the empirically observed patterns of convergence and communicative successes (van de Braak, Dingemanse, Toni, van Rooij, & Blokpoel, 2021). Third, models that seem to overcome the limitations in the former type of models do so by using what we consider a theoretical 'hack', i.e., assuming that the receiver can always disambiguate their inferred referent by pointing at it (a.k.a. *ostension*; "Is that 🍎?") rather than making a potentially ambiguous request to try to get more information ("Is that the prickly one?"). While ostension is sometimes possible and used in naturalistic communication, it is neither always necessary nor always available as an option.[3] Still, people can communicatively align even in those cases.[4]

Uncovering these problems in state-of-the-art models must be seen as a positive instance of the *theoretical problem-finding* paradigm (Adolfi, van de Braak, & Woensdregt, 2023). The value of this type of theoretical problem-finding cannot be overstated. *"[T]heoretical problem-finding allows us to pinpoint where our understanding is lacking [...]"*(Adolfi et al., 2023). While most of cognitive science focuses on empirical problem-solving, there is at present no comparable attention to theoretical problems, leaving the study of cognitive capacities sometimes theoretically underinformed. Therefore it is important to lay out the full landscape of theoretical problems relevant to a phenomenon before trying to tackle any one of them individually. Failing to do this risks us fooling ourselves into thinking we have solved all theoretical problems for this phenomenon, while some still may lurk in the dark. Those hidden, invisible problems can lead to the false impression that full explanations for cognitive capacities have been found.

In this paper, we contribute to theoretical problem-finding in accounts of communicative alignment by identifying another, complementary theoretical obstacle in models of communicative alignment: *intractability*. While points (1)–(3) are all different problems, they all fall under the umbrella of *cognitive scope violation* (as defined by Adolfi et al., 2023). That is, models that experience these problems *undergeneralize* the cognitive capacity. In contrast, *intractability* is a problem experienced by models that *overgeneralize* the cognitive capacity (van Rooij, 2008; Blokpoel, 2018).

Intractable models assume computational resources that grow excessively fast (e.g. exponentially) as a function of the input size. This makes them computationally implausible for all but the smallest toy scenarios (van Rooij, Evans,

Müller, Gedge, & Wareham, 2008; van Rooij, Blokpoel, Kwisthout, & Wareham, 2019). Take as an example the fruit platter scenario (Figure 1). Given an input with a 'vocabulary' of 10 words and a 'context' of 10 fruits, an intractable model postulates computations that require $2^{10 \times 10} = 1.267.650.600.228.229.401.496.703.205.376$ basic computation steps. Even assuming that a human can perform a hypothetical 500 quadrillion computations per millisecond, it would still take $2.535.301.200$ seconds (or about 80,3 years) to compute the intended fruit. A far cry from the time scales realized in natural conversation (Stivers et al., 2009). Even when one can specify the necessary and sufficient computations, they are seemingly inevitably computationally intractable. As such, such a model cannot explain the speed of *communicative alignment* which greatly undermines its explanatory value.

The remainder of this paper is organized as follows. First we present a formalization of an existing model of communicative alignment that includes *ostension*—that is, including the theoretical 'hack' mentioned earlier (van de Braak et al., 2021). Second, we present a mathematical proof of intractability, with full mathematical proof details in the Appendix[5], and explain how this also implies intractability of a model extension without the 'hack'. Hence, we can conclude both types of model embody this same theoretical problem. Subsequently, we present robustness checks for the intractability results and draw out their implications. We close by reflecting on the general challenge that these results pose and how they shine a light on what current explanations of communicative alignment cannot yet explain. This illumination of the problem landscape provides clear theoretical goals for theorists of communication. Intractability is present even in models with theoretical 'hacks', showing how important this investigation of the problem landscape really is.

Computational-level model of communicative alignment

We analyze a computational-level (Marr, 1982) model of communicative alignment (van de Braak et al., 2021). It builds on Rational Speech Act (RSA) models (Frank & Goodman, 2012; Hawkins et al., 2017). RSA models assume that pragmatic inference (i.e., the ability to disambiguate the meaning of utterances) is central to referential communication. Consider the referents (things that can be talked about) 🍐and 🍎, and the utterances *pink* and *prickly*. When interpreting *pink*, people have the capacity to understand that it refers to 🍐 despite the utterance being ambiguous—both fruits can be considered pink.[6] This basic principle, however, presupposes that people share a common understanding of what the utterances can mean: a shared *lexicon* that encodes which utterances refer to which referents (see Table 1).

In natural communication this full lexical alignment is gen-

---

inferences do not involve an interaction as portrayed in the example, but instead take the form: "I think it's 🍎!"

[3]E.g. ostension is not possible when talking on the telephone, when talking about objects that are not physically present and/or when communicating about abstract concepts.

[4]Other models, e.g. Steels (2015), often exclude these cases and assume objects are always available to ostensively point at.

[5]The appendix is available at `https://osf.io/xe9bn`

[6]How? Because if 🍎was the intended referent, one would have said *prickly* and not *pink*.

Box 1: Communicative alignment

To illustrate the phenomenon of (non-ostensive) communicative alignment, consider this collaborative game (adapted from Eijk et al., 2022). One player secretly chooses one the 8 Fribbles. Then both players work together so that the other player can infer which Fribble was secretly chosen. Players can communicate freely, but they cannot point at the Fribbles. To experience communicative alignment we recommend playing this game with another person or, alternatively, read the dialogue and try to discover the secret Fribble.
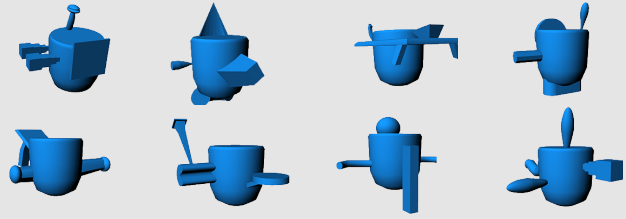
Player 1: *"The Fribble kind-of looks like a waiter."*
Player 2: *"Uh, does it have a round tray?"*
Player 1: *"No, not the one with the trumpet. It has a square tray with legs and a small dish on top."*

	🐉	🔴
prickly	no	yes
pink	yes	yes

Table 1: A simple lexicon encodes utterance-referent relations.

erally not available and needs to be built through interaction (Hutchins & Hazlehurst, 1995; Hayashi, Raymond, & Sidnell, 2013; Kitzinger, 2012; Dingemanse et al., 2023) before mutual understanding can be achieved (Box 1). In the model by van de Braak et al. (2021) two communicating agents allow for the possibility that each utterance can, for their interlocutor, refer to any of the referents. This enables the agents to flexibly adapt and align (Eijk et al., 2022; Kempson, Chatzikyriakidis, & Howes, 2017; Rasenberg, Özyürek, & Dingemanse, 2020). This flexibility is modeled by agents that infer the meaning of an utterance relative to all possible lexica $\mathcal{L}$ (Table 2). This inference combines an agent's preference for particular meanings, modeled as a prior over all possible lexica $\Pr(\mathcal{L})$, with the interactional history h, leading to a posterior distribution over all possible lexica $\Pr(\mathcal{L} \mid h)$.

	🐉	🔴		🐉	🔴		🐉	🔴	
...	*pi*	yes	no	*pi*	yes	no	*pi*	yes	yes	...
	pr	no	yes	*pr*	yes	yes	*pr*	yes	no	

Table 2: Excerpt from all $(2 \times 2)^2 = 16$ simple lexica (*pi* for *pink* and *pr* for *prickly*).

The posterior probability of each possible lexicon $\Pr(\mathcal{L}|h)$ is the model's main computational principle (Appendix A):

POSTERIOR LEXICON DISTRIBUTION (semi-formal)
Input: Prior (preference) over all possible lexica $\Pr(\mathcal{L})$ and the interactional history h.
Output: The posterior probability of each possible lexicon $\Pr(\mathcal{L} \mid h)$, given that utterances and referents are pragmatically inferred.

This computation is the focus of the complexity analysis and it is part of all main model sub-components. The inferences

in each of these components are relative to (the probability of) each possible lexicon:

1. Producing an utterance given an intended referent (INFER DISTRIBUTION OVER SIGNAL and SIGNAL SAMPLER[7])

2. Inferring a referent given an utterance (INFER DISTRIBUTION OVER REFERENT and REFERENT SAMPLER)

3. Inferring if the agent believes they have achieved mutual understanding (PERCEIVED UNDERSTANDING).

See Figure 2 for an illustration of the model and the relationships between the model components.

We next introduce the PERCEIVED UNDERSTANDING computation. It will serve as an illustration for the complexity proofs. PERCEIVED UNDERSTANDING characterizes how, given an observed utterance u, agents decide if they perceive their inference to be certain. If many referents are probable then certainty is low and, vice versa, if only a few are probable then certainty is high. With low certainty, agents decide to continue the conversation (red arrows, Fig. 2). The inference is relative to the history h and all possible lexica $\mathcal{L}$ and it uses POSTERIOR LEXICON DISTRIBUTION $\Pr(\mathcal{L} \mid h)$ as a sub-computation.

PERCEIVED UNDERSTANDING (semi-formal)
Input: Prior over all possible lexica $\Pr(\mathcal{L})$, the history h, an observed utterance u, and a certainty threshold η.
Output: Is the certainty of the inferred referent of u high? That is, is the entropy $H(\Pr(r \mid u, h)) < \eta$? Here, the probability over referents is relative to all possible lexica and the history, and the distribution over all possible lexica is as defined in POSTERIOR LEXICON DISTRIBUTION:

$$\Pr(r \mid u, h) = \sum_{\mathcal{L}} \Pr(r \mid u)\Pr(\mathcal{L} \mid h)$$

Earlier, we introduced the notion of ostensive communication: unambiguous communication through e.g. explicit pointing. Van de Braak et al. (2021) have proposed two models for communicative alignment. One model assumes ostension, the other does not. The difference between the models lies in the history h of each turn t_i. Ostensive history stores the initiator's utterance (e.g., 'dragon fruit') and

---

[7]These components are named with the more general term signal.
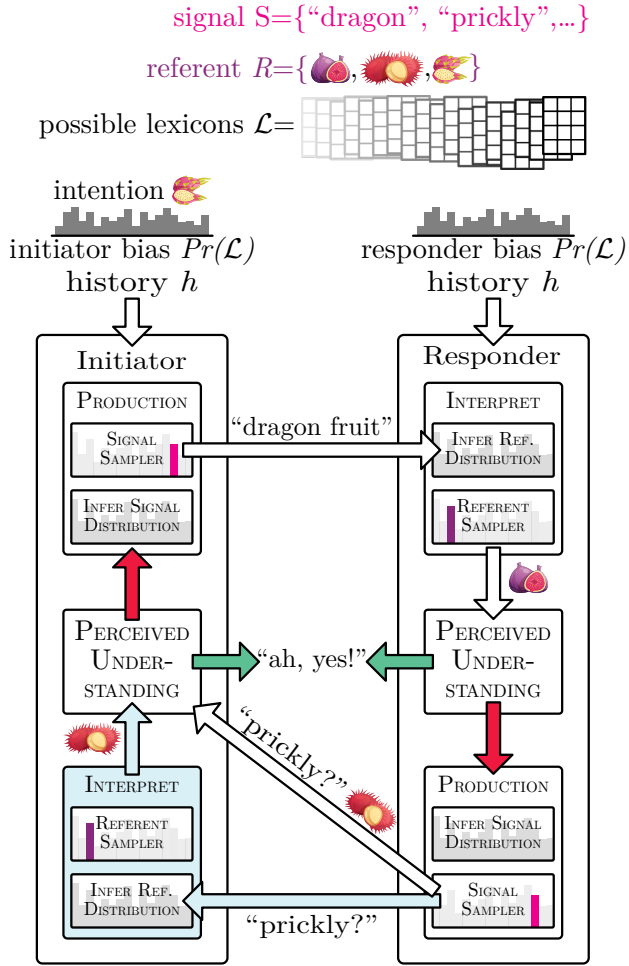
Figure 2: An overview of the computational model. Arrows denote input-output relations. Note the two model variants, ostensive and non-ostensive, where the non-ostensive model uses the INTERPRET SIGNAL sub-component because the referent is not explicitly and unambiguously given (blue arrows). Red arrows indicate perceived non-understanding, the green show perceived understanding.

their inferred referent (🥭). Non-ostensive history stores the initiator's utterance (e.g., 'dragon fruit') and the responder's repair request (e.g., 'prickly?') (see also Fig. 2).

Consequently, the computation underlying $Pr(\mathcal{L} \mid h)$ in POSTERIOR LEXICON DISTRIBUTION is different. The non-ostensive agents require additional inference to interpret the responder's repair request, which is given for free in the ostensive model (see Fig. 2, and Appendix A.2).

Computational complexity analysis

First, we briefly introduce complexity-theoretic concepts and techniques at an conceptual level after which we illustrate intractability of PERCEIVED UNDERSTANDING. Then, we cover results for all components of the model, for which intractability proofs are derived from the intractability of PERCEIVED UNDERSTANDING. For more formal details, we kindly refer the reader to Appendix B.1 and articles and text-

books (Garey & Johnson, 1979; van Rooij et al., 2008; Sanjeev & Barak, 2009; van Rooij et al., 2019).

Intractability proof

We prove intractability for PERCEIVED UNDERSTANDING (PU) by *reduction* from a known intractable problem, viz. 3-satisfiability (3SAT), in two steps:

1. Prove that a tractable transformation algorithm A exists from any 3SAT input to a PU input.

2. Prove that the output given by PU is the correct output for 3SAT.

Given the reduction constituted by (1) and (2), PU must be intractable. This follows by contradiction (Fig. 3). Suppose that PERCEIVED UNDERSTANDING could be solved tractably. Then one could compute 3SAT tractably, viz. via the 'shortcut' through algorithm A and PU. This is impossible since 3SAT is intractable, hence PU must be intractable.
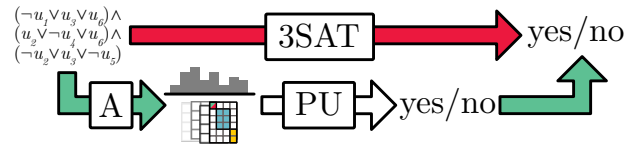
Figure 3: PERCEIVED UNDERSTANDING cannot be tractable (green), as this would contradict 3SAT's intractability (red).

Key to the proof is the algorithm A. To illustrate its design, we first define 3SAT (from van Rooij et al., 2019):

3SAT
Input: A set of Boolean variables $U = \{u_1, \ldots, u_n\}$ and a set of clauses $C = \{c_1, \ldots, c_m\}$, where each clause has exactly three literals.
Output: Is there a truth assignment $t : U \to \{true, false\}$ such that all clauses C are satisfied?

This computational problem takes as input a logical predicate of any length, e.g., $(u_1 \lor u_2 \lor \neg u_3) \land (\neg u_2 \lor u_3 \lor u_5)$, and outputs 'yes' if there exists a truth value assignment to all literals u_i such that the predicate is true, e.g., $(T \lor F \lor \neg T) \land (\neg F \lor T \lor F) = T$.

Transformation A takes any 3SAT input and creates a matching input for PERCEIVED UNDERSTANDING, such that PU outputs an answer for 3SAT. The transformation exploits the fact that all possible lexica represent all possible binary strings. These strings can encode all possible 3SAT candidate solutions. This prior probability function $Pr(\mathcal{L})$ is constructed to give a non-zero probability to a lexicon $L \in \mathcal{L}$ if it encodes the 3SAT instance, and if the candidate solution is correct (Fig. 4). This setup ensures that PU outputs yes if and only if 3SAT would output yes, proving intractability.

Results

In this section we present an overview of the complexity-theoretic results. For full mathematical details and formal
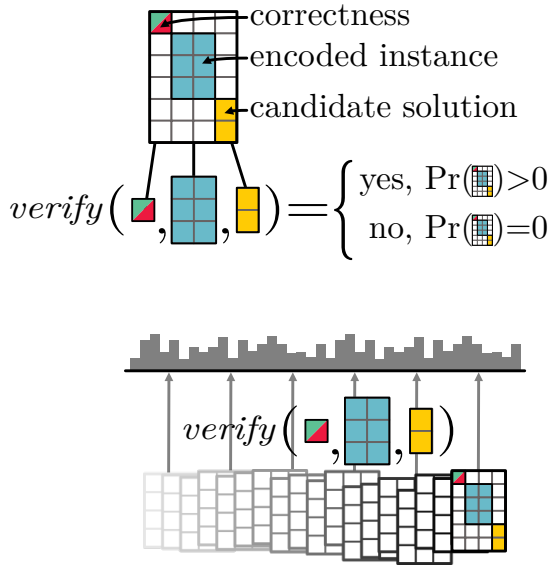
Figure 4: An illustration of the transformation algorithm A. The prior probability over all possible lexica evaluates encoded candidate 3SAT solutions.

proofs, we refer to the reader to Appendix B. Within the Ostensive Communication model, the following results apply:

Result 1 PERCEIVED UNDERSTANDING *is intractable (Appendix B.2).*

Result 2 INFER DISTRIBUTION OVER REFERENTS *and* INFER DISTRIBUTION OVER SIGNALS *are intractable (Appendix B.3).*

Result 3 REFERENT SAMPLER *and* SIGNAL SAMPLER *are intractable (Appendix B.4).*

Furthermore, Results 1, 2 and 3 hold under a wide range of conditions that seemingly simplify the computational model, yet do not remedy intractability. This emphasizes that the intractability is a foundational property of the model and not an artifact. Table 3 shows this robustness of our results.

Model component	history h	entropy η	order n	prior $\Pr(\mathcal{L})$
PERCEIVED UNDERSTANDING	any	1	≥ 0	$O(n^c)$
PERCEIVED UNDERSTANDING	$|h| \geq 1$	$0 < \eta$	≥ 0	$O(n^c)$
INFER DISTRIBUTION OVER REFERENTS	any	–	≥ 0	$O(n^c)$
INFER DISTRIBUTION OVER SIGNALS	any	–	≥ 0	$O(n^c)$
REFERENT SAMPLER	$|h| = 1$	–	≥ 0	$O(n^c)$
SIGNAL SAMPLER	$|h| = 1$	–	≥ 0	$O(n^c)$

Table 3: Robustness of the intractability results for the main parameters of each model component.

Intractability Results 1, 2, and 3 generalize to the Non-ostensive Communication model under the same conditions as listed in results 1, 2, and 3.

Corollary 1 PERCEIVED UNDERSTANDING, INFER DISTRIBUTION OVER REFERENTS, INFER DISTRIBUTION OVER SIGNALS, REFERENT SAMPLER *and* SIGNAL SAMPLER *in the Non-ostensive Communication model are intractable (Appendix B.6).*

Interpretation of results

Results 1, 2, and 3 show that the Ostensive Communication model is intractable, implying that it cannot explain the speed of communicative alignment in the real world. Furthermore, the intractability is not contained in a single sub-component of the model. Rather, all main sub-components of the model are intractable (see Figure 2). To explain how communicative alignment can occur on realistic time scales one would have to revise all sub-components.

Any such revision cannot be simple. In addition to the intractability results, Table 3 illustrates that the intractability is robust relative to a range of parameters, conditions and variations. Parameter robustness implies that one cannot make any assumption on the parameters within the specified bounds that would render the model computationally tractable.

History h. Results 1, 2, and 3 are robust for the history h. Specifically, the sub-computations are intractable when h contains at most one element, implying that within this model no realistic assumption on the content of h can make the sub-components tractable. More information in the history is not going to make the intractability go away.

Entropy threshold η. Entropy characterizes the level of certainty that agents need to perceive understanding. Result 1 is robust for an entropy threshold $\eta > 0$, assuming $|h| \geq 1$. Specifically, PERCEIVED UNDERSTANDING is intractable for $\eta = 1$ and also intractable for all other $\eta > 0$ given there has been at least a single turn in the history. Allowing for arbitrary levels of uncertainty is not going to make the intractability go away.

Order of reasoning n. Results 1, 2, and 3 hold for any order of reasoning $n \geq 0$. Even when agents do no pragmatic reasoning (i.e., $n = 0$) the model is intractable (see also van de Pol, van Rooij, & Szymanik, 2018) .

Prior $\Pr(\mathcal{L})$. Results 1, 2, and 3 assume a polynomial-time computable prior function. Thus, the model's intractability is not due to an intractable prior function, but core to the model itself.

Sampling. Result 3 implies that optimal inference is not causing the model's intractability, as it is not possible to use (non-optimal) sampling as a method to get around the intractability.

Non-ostensive Communication. Corollary 1 indicates that Results 1, 2 and 3 and the conditions also hold for the Non-ostensive Communication model.

Discussion

How people are able to understand each other even when talking about novel things and without a fully shared vocabulary (what we call *communicative alignment*) is challenging to explain computationally. So far, (1) empirically corroborated models only work under shared knowledge and vocabularies, and leave out interactive processes needed to overcome misalignment; (2) models that do include misalignment and interactive processes cannot account for communicative successes in real-world conditions; (3) models that overcome the limits in (2) use a theoretical 'hack'. With our analyses we have added an extra theoretical challenge: models of type (2) and (3) hide computational intractability in all their core components (see Figure 2).

Our intractability results imply that so far these models make unrealistic assumptions about the computational resources available to real-world, embodied communicators, who need to operate under limited resources. Here we explain why this problem is not a mere technicality that can be brushed off, but a deep theoretical challenge that cannot be easily dissolved. We remind the reader that this is a strength of this theoretical problem: it can be robustly used to improve our understanding of what we do not yet scientifically understand (Adolfi et al., 2023) about communicative alignment and help carve/sculpt better, more explanatory theories (Blokpoel, 2018).

To see why the proven intractability is not a mere technical quirk nor easy to 'hack' out of models, note that the core intractability result holds for a wide range of parameters, conditions, and variations of the models (see **Interpretation of Results**). Moreover, the intractability cannot be removed by common appeals to approximability heuristics or even 'as if' explanations (van Rooij, Wright, Kwisthout, & Wareham, 2018). It is known that claims to approximability generally run into intractability as well, because intractable problems are not standardly approximable. Even when they are, the fall short of criteria of approximation needed for cognitive science explanations (van Rooij & Wareham, 2012). Accordingly, we have an inapproximability proof that show that the probabilistic inferences postulated in the models cannot be approximated by efficient sampling (Appendix B.4; (cf. Sanborn, 2017; Kwisthout, 2018)). While heuristics generally are tractable by design, they necessarily cannot have any guarantees for approximating with the computations postulated at the computational level (van Rooij, Wright, & Wareham, 2012). In other words, heuristics at the algorithmic level can only 'buy' tractability by 'cutting away' parts of the phenomenon as modeled at the computational level, meaning we still fail to explain the original phenomenon.[8]

Trying to cope with intractability at the computational level by 'cutting away' parts of the phenomenon is a more common strategy in the literature. These routes, unfortunately, lead at best to an impoverished understanding of the phenomenon. For instance, modeling 'by the slice' of communication can be seen as an instance of buying tractability by distorting or simplifying the phenomenon (e.g., models of type (1) mentioned earlier). Another is to assume ostension to always be available (Hawkins et al., 2017), see also the Ostensive model (model type (3)). Such approaches or 'hacks' yield a situation where the phenomenon is oversimplified and the model conditions are overconstrained relative to the real world situations. Probably contrary to intuition, the ostension 'hack' does not even buy tractability (see Result 1). This does not yet rule out that other intuitions may help yield tractability at the computational level. We consider two more candidates.

First, we consider the intuition that interaction makes it easier to have successful communication (van Arkel, Woensdregt, Dingemanse, & Blokpoel, 2020; Dingemanse, 2020; A. Clark, 2006; Risko & Gilbert, 2016). At first glance, this intuition may seem to contradict our proof results. However, this reflects a map-territory confusion (cf. Guest & Martin, 2023; Guest, 2024). Namely, the intuition relates to the real-world *phenomenon* ('territory'), whereas intractability is a property of a theoretical *model* ('map'). Any (non-formal) intuition requires further computational investigation. This would entail defining what 'easier' means, as it could refer to either fewer computational resources or higher probability for successful communication. As it stands, the intractability proof we presented implies that *the model* cannot explain how interaction could lead to a reduction in computational resources as intuited to be the case for the *phenomenon*. In fact, it is exactly by trying to explain how people can communicatively align through interaction that intractability is revealed in our models which emphasizes the theoretical challenge of computationally explaining the intuition.

Second, we consider the intuition that in daily communication communicative alignment is easier because interlocutors rarely, if ever, converge on overlapping lexica (Stolk, Bašnáková, & Toni, 2022). The claim implicitly assumes that the model requires agents to converge on overlapping lexica. This is not the case, as agents in the model make decisions based on their *perceived* understanding rather than on any factual convergence. Hence, full (or even partial) lexical convergence cannot be responsible for the intractability result. Furthermore, an incorrect assumption in this intuition is that the intractability only arises from the full interaction (presumably required for convergence) and thus that eliminating the necessity for lexical convergence will make the interaction computationally easier. However, the intractability is present in every agent-level model component (see Results 1-3 and Figure 2), even before any interaction occurs. Thus, incorporating this intuition in the model would not resolve the intractability.

Of course, the above considerations do not exhaust all possible intuitions one may have about what would or would not yield tractability at the computational level of explana-

---

[8]While systematic 'cutting' could be done in a principled way (using parameterized complexity analyses; see (van Rooij & Wareham, 2007; van Rooij et al., 2019)), our results so far do not yet yield constraints that can make the computational level model tractable.

tion. We do hope that our discussion of them illustrates that such intuitions often involve a confusion between explanation (model) and explanandum (phenomenon), and how one can oneself catch this confusion. Also, we hope to have illustrated that even if the intuitions are correct those intuitions may not translate directly or rigorously, for instance because they are underspecified. Further, even when these intuitions could be translated into the model, they still may not yield tractability. Theoreticians who think they have identified a solution to this intractability problem should be self-critical. Verifying whether or not a model change yields (in)tractability requires formal modeling and complexity-theoretic proof methods such as we have adopted in this paper.

We close by reflecting on implications of these results for computational explanations of communicative alignment. As we said from the start, theoretical problem finding is useful as it shines a light on gaps in our scientific understanding. Clearly, we do not yet scientifically understand how people can communicatively align when lacking a fully shared vocabulary. It is exactly by trying to figure out ways in which these gaps can and cannot be filled that we advance our scientific understanding, and progressively sculpt better, more explanatory theories. Importantly, this process of filling the gaps and sculpting better, more explanatory theories should not be thought of as a recipe and there are no theoretical 'quick fixes' (Devezer, 2024; Rich, de Haan, Wareham, & van Rooij, 2021). We invite theorists to take on the theoretical challenges and let them inform development of future theories of communicative alignment.

Acknowledgments

References

Adolfi, F. G., van de Braak, L. D., & Woensdregt, M. (2023). *From empirical problem-solving to theoretical problem-finding perspectives on the cognitive sciences*. PsyArXiv.

van Arkel, J., Woensdregt, M., Dingemanse, M., & Blokpoel, M. (2020). A simple repair mechanism can alleviate computational demands of pragmatic reasoning: simulations and complexity analysis. *PsyArXiv*.

Barry, T., Griffith, J., De Rossi, S., & Hermans, D. (2014). Meet the Fribbles: novel stimuli for use within behavioural research. *Frontiers in Psychology*, *5*.

Blokpoel, M. (2018). Sculpting Computational-Level Models. *Topics in Cognitive Science*, *10*(3), 641–648.

van de Braak, L. D., Dingemanse, M., Toni, I., van Rooij, I., & Blokpoel, M. (2021). Computational challenges in explaining communication: How deep the rabbit hole goes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Clark, A. (2006). Material Symbols. *Philosophical Psychology*, *19*(3), 291–307.

Clark, H. H., & Schaefer, E. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, *2*(1), 19–41.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Devezer, B. (2024). There are no shortcuts to theory. *Behavioral and Brain Sciences*, *47*, e38.

Dingemanse, M. (2020). Resource-rationality beyond individual minds: the case of interactive language use. *Behavioral and Brain Sciences*, *43*, e9. (Publisher: Cambridge University Press)

Dingemanse, M., Liesenfeld, A., Rasenberg, M., Albert, S., Ameka, F. K., Birhane, A., ... Wiltschko, M. (2023). Beyond Single-Mindedness: A Figure-Ground Reversal for the Cognitive Sciences. *Cognitive Science*, *47*(1), e13230.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., ... Enfield, N. J. (2015). Universal Principles in the Repair of Communication Problems. *PLOS ONE*, *10*(9), e0136100.

Dynel, M. (2020). To say the least: Where deceptively withholding information ends and lying begins. *Topics in Cognitive Science*, *12*(2), 555–582.

Eijk, L., Rasenberg, M., Arnese, F., Blokpoel, M., Dingemanse, M., Doeller, C. F., ... Bögels, S. (2022). The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, *264*, 119734.

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. San Francisco, CA: W. H. Freeman.

Guest, O. (2024). What Makes a Good Theory, and How Do We Make a Theory Good? *Computational Brain & Behavior*.

Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Com-

putational Brain & Behavior, 6(2), 213–227.

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (p. 6). Austin, TX: Cognitive Science Society.

Hayashi, M., Raymond, G., & Sidnell, J. (Eds.). (2013). *Conversational Repair and Human Understanding*. Cambridge: Cambridge University Press.

Holler, J., & Wilkin, K. (2011). Co-Speech Gesture Mimicry in the Process of Collaborative Referring During Face-to-Face Dialogue. *Journal of Nonverbal Behavior*, 35(2), 133–153.

Hutchins, E., & Hazlehurst, B. (1995). How to invent a shared lexicon: the emergence of shared form-meaning mappings in interaction. In *Social Intelligence and Interaction: Expressions and implications of the social bias in human intelligence* (pp. 53–67). Cambridge University Press.

Kempson, R., Chatzikyriakidis, S., & Howes, C. (2017, July). Cognitive science, language as a tool for interaction, and a new look at language evolution..

Kitzinger, C. (2012). Repair. In *The Handbook of Conversation Analysis* (pp. 229–256). John Wiley & Sons, Ltd.

Kwisthout, J. (2018). Approximate inference in Bayesian networks: Parameterized complexity results. *International Journal of Approximate Reasoning*, 93, 119–131.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

Pohlhaus, G. (2012). Relational knowing and epistemic injustice: Toward a theory of willful hermeneutical ignorance. *Hypatia*, 27(4), 715–735.

van de Pol, I., van Rooij, I., & Szymanik, J. (2018). Parameterized Complexity of Theory of Mind Reasoning in Dynamic Epistemic Logic. *Journal of Logic, Language and Information*, 27(3), 255–294.

Quine, W. V. O. (2013). *Word and Object* (Illustrated edition ed.). Martino Fine Books.

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in Multimodal Interaction: An Integrative Framework. *Cognitive Science*, 44(11), e12911.

Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in cognitive sciences*, 20(9), 676–688.

van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science: A Multidisciplinary Journal*, 32(6), 939–984.

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis*. Cambridge University Press.

van Rooij, I., Evans, P., Müller, M., Gedge, J., & Wareham, T. (2008). Identifying Sources of Intractability in Cog-

nitive Models: An Illustration using Analogical Structure Mapping. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 30(30).

van Rooij, I., & Wareham, T. (2007). Parameterized Complexity in Cognitive Modeling: Foundations, Applications and Opportunities. *The Computer Journal*, 51(3), 385–404.

van Rooij, I., & Wareham, T. (2012). Intractability and approximation of optimization theories of cognition. *Journal of Mathematical Psychology*, 56, 232–247.

van Rooij, I., Wright, C., & Wareham, T. (2012). Intractability and the use of heuristics in psychological explanations. *Synthese*, 187, 471–487.

van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, 195(2), 491–510.

Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, 112, 98–101.

Sanjeev, A., & Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge ; New York: Cambridge University Press.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382.

Steels, L. (2015). *The Talking Heads experiment: Origins of words and meanings*. Language Science Press.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., . . . Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.

Stolk, A., Bašnáková, J., & Toni, I. (2022). Joint epistemic engineering: The neglected process in human communication. In *The Routledge Handbook of Semiosis and the Brain.* Routledge.

Predicting Complex Problem Solving Performance in the Tailorshop Scenario

Daniel Brand (daniel.brand@metech.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Sara Todorovikj (sara.todorovikj@metech.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Abstract

Complex problem solving (CPS) is a fundamental capability of humans. It is often studied through microworlds, with the Tailorshop-scenario as a well-investigated prominent example. This paper addresses several research questions for CPS in the Tailorshop scenario: Firstly, it examines the impact of background knowledge vs. understanding underlying dynamics. Secondly, it investigates the predictability of a participants' performance, particularly when considering their assumptions about the scenario's mechanisms. Finally, it discusses the suitability of the Tailorshop as a scenario for cognitive modeling of CPS. Thereby, we discuss some of the measures that have been proposed to assess CPS performance, considering CPS from an perspective of predictive modeling. Based on our results, we conclude that effective prediction of outcomes in complex tasks necessitates uniform impact of actions throughout, facilitating comprehension of both overarching strategies and smaller adjustments crucial in real-world problem-solving domains.

Keywords: Complex Problem Solving; Causal Map; Mental Representations; Cognitive Modeling; Tailorshop

Introduction

In our everyday life, individuals regularly encounter complex systems spanning societal, economic, and environmental realms with many latent variables, requiring adept problem-solving and decision-making skills. However, traditional decision-making research often occurs in small controlled settings, raising concerns about its relevance to real-world complexities (Pitz & Sachs, 1984). To address this, complex dynamic tasks, known as dynamic decision-making (DDM), have been used to study Complex Problem Solving (CPS) behavior. DDM involves participants making decisions within dynamic environments, observed as outcomes that may or may not be affected by decisions made (Edwards, 1962). Computer simulations, called *microworlds*, provide realistic environments for studying complex problem-solving and decision-making processes. These studies challenge cognitive demands regarding goal elaboration, information search, hypothesis formation and forecasting, which ultimately rely on an individual's planning and decision making capabilities, but also creativity (Dörner & Wearing, 1995; Gonzalez, Vanyukov, & Martin, 2005; Funke, 2014). The microworld Tailorshop (e.g., Putz-Osterloh, 1981, 1983; Funke, 1988; Danner et al., 2011; Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015) is an extensively studied computer-based dynamic decision-making scenario for CPS. Participants assume the role of a tailorshop manager for 12 months, tasked with purchasing raw materials, managing production capacity, and maximizing profit by selling shirts. The environment comprises 24 variables, with 21 visible to participants and 12 directly manipulable. These variables are interconnected, with modifications to one potentially impacting others in subsequent simulated months (e.g., advertising influences customer interest, which then affects sales). Tailorshop has been utilized to explore problem-solving processes, intelligence, and professional performance among others (Danner et al., 2011). Success in Tailorshop is typically defined as a consistent increase in company value over months, with the first month excluded from scoring to enhance consistency with the 2-12 months score being a reliable predictor for success, as found by previous studies (Danner et al., 2011; Greiff et al., 2015). However, Greiff and Funke (2009) criticize the "one-item-testing" of one large, complicated scenario as a severe shortcoming of CPS research. They propose that the detection of individual differences could be facilitated by a formal framework of linear structural equation systems — the MicroDYN approach. Instead of a single, complex system, subjects engage with 8-12 items to explore, detect causal relations between variables, draw connections between them to represent their mental model, and then adjust values to achieve target outcomes.

In the light of the discussion in the current state of the art, this paper presents a rigorous analysis of the Tailorshop scenario from a predictive modeling perspective: (1) Investigating how prior knowledge and individual characteristics influences behavior and assess their worth as a predictor; (2) Search for action patterns that can serve as a base for modeling endeavours; and (3) discuss the predictability of participants' performance and the suitability of the Tailorshop scenario for predictive modeling of CPS as a whole. Thereby, the structure of the paper is as follows: the next section presents the experimental data, followed by an introduction to the causal map analysis in Section 3. Section 4 outlines initial implications drawn from our analyses and tests the relationship between causal map information, strategies, participant actions, and performance. Finally, a discussion addressing the aforementioned key issues concludes the paper.

Experiment

Participants and Materials. We conducted an onsite study in German in our lab involving 52 students at the Chemnitz
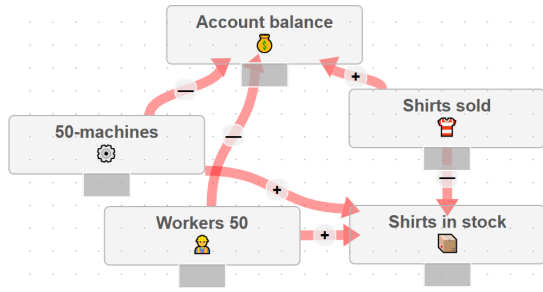
Figure 1: Illustrative example for a causal map created in the graphical user interface used by the participants to represent the relationships between Tailorshop variables (cp. Table 1).

University of Technology. Participants were compensated with either course credits or monetary rewards. The Tailorshop simulation was based on the implementation by Danner et al. (2011)[1]. Similar to the drawing of variable connections in MicroDyn (e.g., Greiff & Funke, 2009), we aimed at obtaining information about the understanding of the relationships between variables (cp. Table 1) in the scenario. Therefore, we developed a graphical interface that allowed participants to represent their understanding in the form of a causal map (Figure 1 shows an illustrative example).

Procedure. Prior to the Tailorshop experiment, participants completed the German version of the Need for Cognition questionnaire (NFC; Beißert, Köhler, Rempel, & Beierlein, 2015) and a 7-question version of the Cognitive Reflection Task (Toplak, West, & Stanovich, 2014). They were then introduced to the Tailorshop topic without explaining any of its mechanisms. Subsequently, participants were presented with variables within the causal map tool and asked to delineate connections denoting relationships between them labelling these connections as positive or negative. Afterwards, participants had an exploration phase of 6 simulated months with the Tailorshop simulator. Following the exploration phase, the scenario was reset, and participants performed a 12-month testing phase. Post testing, participants were asked to construct another causal map to assess their comprehension. Then they were asked for their specific strategies and rated variable relevance using a 5-point Likert scale. All collected data and associated scripts are publicly accessible on GitHub[2].

Analyzing the Causal Maps

For the analysis 4 participants had to be excluded, since they skipped a causal map, leading to a dataset containing the responses of 48 participants (30 female, 17 male, 1 diverse).

Causal Map Properties

Figure 2 shows the aggregated graphs from participants' causal maps both before and after engaging with the Tailor-

---

[1] https://www.psychologie.uni-heidelberg.de/ae/allg/tools/tailorshop/index.html

[2] https://github.com/brand-d/iccm2024-tailorshop

shop simulation. Only edges reported by at least 5% of participants are depicted. Additionally, this figure includes relationships derived from the Tailorshop implementation for comparative analysis. For simplicity, the graphs representing the causal maps of participants before and after interaction with the Tailorshop will be referred to as *Before* and *After* for the remainder of this paper.

Variables controllable by participants, denoted in lightblue, are intentionally designed to be not influenced by other variables within the Tailorshop – in contrast to potential interconnections in the real world. Therefore, edges towards these variables are represented as dotted lines in the graph. This presentation form aims to highlight other edges directly comparable to the Tailorshop simulator.

The core discrepancy is between the Tailorshop graph and participants' causal maps. The simulator graph demonstrates mostly direct connections to few key variables such as account, shirt, and material stock. However, participants' causal maps exhibit higher levels of indirection and interconnection: For instance, while workers are not directly linked to costs, they are indirectly influenced by factors like salary, even when denoted on a per-person basis. Moreover, the connections in participants' causal maps also cover real-world connections that go beyond the scope of the simulation. For instance, the influence of location on worker satisfaction is identified, a *soft factor* relationship not covered by the simulator. While the first differences are most likely caused by a less formal understanding of the concepts, the latter is an expected problem of a real-world based simulation, since a simulation will automatically fall short in some aspects, which can lead to some false assumptions by the participants. The difference in interconnectivity is also visible with respect to the number of incoming and outgoing edges (see Table 2). The Tailorshop simulation has a few central nodes (e.g., the bank account) where everything comes together, while other nodes have no incoming edges at all (i.e., the variables controlled by the participants), whereas no such extremes are visible in the participants' graphs. The table also shows that the differences between *Before* and *After* are slim, indicating that no substantial structural changes occurred. Although subtle, with some adjustments to the aforementioned problems (i.e., interactions between location and worker satisfaction is no longer present) seemed to have taken place.

In order to quantify the changes between *Before* and *After*, we calculated the similarity between the participants' graphs, and the tailorshop graph. If participants adjusted their assumptions based on experiences with the tailorshop simulation, the changes between *Before* and *After* should lead to an increased similarity with the tailorshop graph. We used the average of the cosine similarities between the adjacency vectors for each node, leading to an overall similarity of .247 for *Before* and .255 for *After*. This change was not significant (Mann-Whitney-U: $U = 1113.5$, $p = 0.781$), which confirms the observation that participants overall did not revise their assumptions to a greater extend.

(a) Before playing the scenario (b) After playing the scenario (c) Tailorshop Simulation
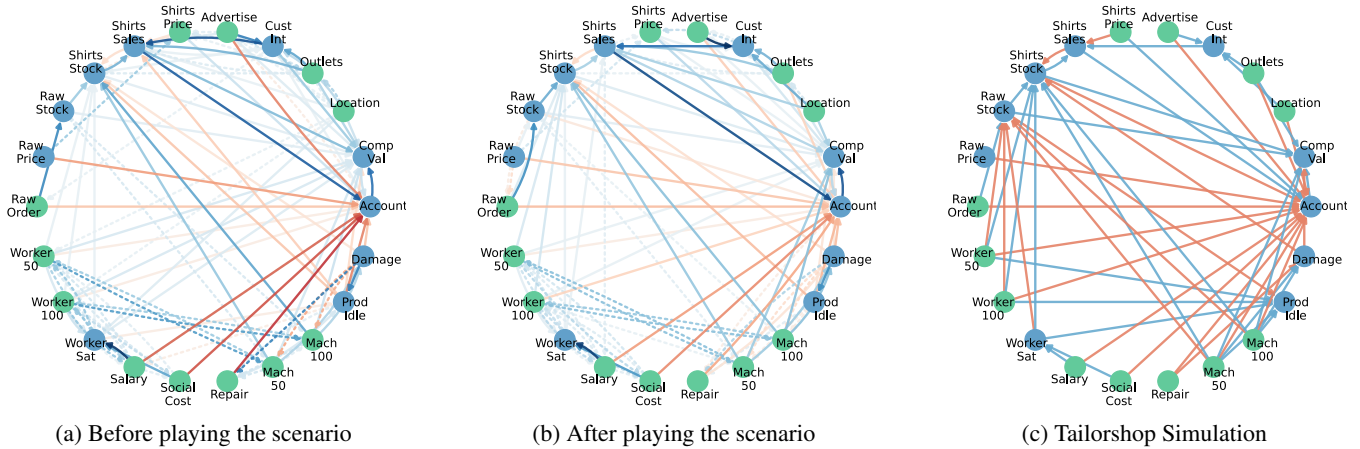
Figure 2: Causal maps before (a) and after (b) playing the tailorshop scenario alongside the graph depicting actual dependencies in the simulation. Blue/Red edges indicate positive/negative relationships, respectively. Darker shades indicate a higher proportion of the respective edge. Green nodes denote controllable variables, while blue nodes represent derived variables. Edges to controllable variables are dotted. Edges reported by less than 5% of participants are omitted.

Table 1: Importance of a variable defined as the average number of paths leading to *Company Value* in the individual *before* and *after* graphs and in the Tailorshop implementation (TS). Average relevance (Rel) of the respective variable reported by the participants is included, variables controllable by participants are excluded.

	Var. Importance			Rel.
	Before	After	TS	
Company Value	(40.58)	(38.35)	(111)	-
Bank Account	24.19	25.31	67	-
Customer Interest	9.06	8.69	9	4.27
Shirts Sales	15.58	15.62	36	4.69
Shirts in Stock	7.19	9.77	72	3.79
Raw Material Price	0.46	1.46	0	3.56
Raw Material Stock	0.96	4.23	32	3.98
Worker Satisfaction	11.0	7.15	14	2.85
Production Idle	2.9	3.83	0	3.30
Damage	3.33	3.1	12	3.25

Table 2: Overview of the graph connectivity comparing the number of incoming and outgoing edges for the graphs from the causal map and the implementation of the tailorshop.

		Mean	SD	Min	Max
	Before	1.46	1.01	0.46	5.19
Incoming	After	1.39	1.18	0.29	5.79
	Tailorshop	2.41	3.77	0	15
	Before	1.46	0.36	0.52	2.31
Outgoing	After	1.39	0.33	0.52	1.96
	Tailorshop	2.41	1.53	0	6

ble 1 shows the importance values for all derived variables (i.e., variables not directly controllable) as well as the relevance that participants provided at the end of the experiment. Note that the bank account was excluded from the relevance, since it was directly explained to be a part of the company value, rendering its relevance trivial. Unsurprisingly given the scenario, shirt sales was assigned the highest relevance, which was also reflected by the importance (15.62 for *After*). Apart from that, no clear correspondence between importance and relevance was visible. However, the importances of the participants' graphs are generally in line with the importances of the variables in the actual tailorshop simulation (Kendall's Tau between *Before* and *TS*: $\tau_b = 0.556$, $p = .029$), indicating that the general concepts are comparable. The relevance, on the other hand, seemed to be mostly focusing on directly sales-related concepts (i.e., shirt sales, and the customer interest), rating variables for production generally lower.

Causal Map and Performance

Since the assumptions participants have about the mechanisms underlying the tailorshop scenario are likely to influence their actions, we investigated the connection between

The graphs obtained from the causal map can also allow for estimates of a variables importance. Since the maximization of the company value was the goal of the tailorshop scenario and participants were instructed to try to do so, we evaluate the importance of a variable with respect to company value. As an importance metric for a variable, we used the number of occurrences in all (cycle-free) paths leading to company value, excluding those starting at the respective variable. Put differently, since edges in the graph denote a positive or negative relationship between variables, the metric gives an estimate of the number of ways a variable influences the company value indirectly when another variable is changed. Ta-

the causal maps and the performance in the tailorshop, assuming that participants with a *Before* graph more similar to the actual tailorshop graph will achieve a better performance. To assess performance, we considered the 11th month as a reference point for the final performance, since participants can skew the results by selling everything in the last month (25% of participants stated that they considered that strategy). Unlike Danner et al. (2011), we use the total difference in company value, since participants were instructed to maximize it until the end of the run (and not consistently each month). We argue that modeling should focus on a task as closely related to the actual instructions as possible. Additionally, to normalize the values of the Tailorshop, we represented the performance as a proportion of the company value change (i.e., by calculating $perf = (cv_{11} - cv_0)/cv_0$, where cv_{11} is the company value at the end of month 11 and cv_0 the initial company value). Subsequently, we proceeded by splitting the participants into two groups based on the median difference in company value between the beginning and the last month. Here, the differences are more apparent: The high performing group had an average similarity of .279 between *Before* and the tailorshop graph, which increased to .319 for *After*. In comparison, the low performing group started with a similarity of .215, which decreased to .191. This indicates that participants that already started out in line with the tailorshops mechanisms were able to further adjust their assumptions, while the low performing group seemed to struggle to grasp the mechanisms. Based on these findings, we aimed to predict the performance in two ways: 1) We used a Support Vector Regression (SVR; for an overview, see Awad & Khanna, 2015) as a simple general-purpose model to predict the performance in the tailorshop for all individual participants based on the *Before* graph, and 2) used the similarity directly as an estimate for the tailorshop performance. First, the SVR was trained and tested using a leave-one-out cross-validation, to ensure that the limited number of participants for a machine learning method is used efficiently. The adjacency matrix of the *Before* graph was used as the input, while the performance value described beforehand was used as the target. The Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2) were used to measure the performance. The median and mean of the target values were added as baseline models, since they represent the optimal constants to minimize MAE and RMSE, respectively.

The results are shown in Table 3. The results show that the SVR was not able to leverage any of the information available in the graph, achieving a similar performance than the mean and median. Including additional individual information (CRT and NFC) did not improve the performance. Two possible explanations for that hinge on fundamental attributes of the present data: First, the coarse structure of the present causal maps only reflect relationships, but do not capture the meaning or importance of certain connections. Second, the tailorshop simulation is a complex, non-linear scenario, that

Table 3: Results of a leave-one-out cross-validation analysis for predicting the tailorshop performance. The table shows the MAE, RMSE and R^2 for the Support Vector Regression (SVR) based on the causal map graph provided before the tailorshop, the SVR based on actions in the first month and the mean and median target value as baseline predictors.

Predictor	MAE	RMSE	R^2
Performance Mean	0.298	0.378	0
Performance Median	0.295	0.381	-0.018
SVR (*Before* graph)	0.293	0.379	-0.007
SVR (First Month Actions)	0.255	0.328	0.247

(a) Profitable Tailorshop group

(b) Unprofitable Tailorshop group
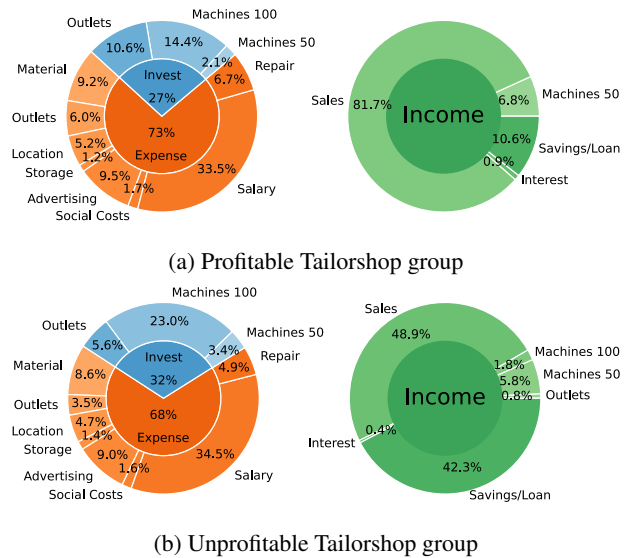
Figure 3: Comparison of the income, expense, and investment proportions of the two participant groups.

can provide greatly differing experiences even for participants with a rather similar overall behavior.

However, even when direct predictions based on the causal map graphs were not possible, the similarity to the tailorshop graph can still serve as a predictor of performance in terms of correlation: If participants started with a graph more similar to the actual simulation, they should be able to make better informed decisions, thereby increasing their performance. The one-sided Spearman correlation between the similarity of the *Before* graph to the tailorshop graph showed a significant moderate correlation ($r = .264$; $p = .035$). Still, it does not seem to provide enough information for the models for individual predictions, but can be a useful utility metric. Similarly, the CRT showed a significant correlation with performance, while the NFC did not (One-sided Spearman rank correlation: CRT: $r = .245$, $p = .047$; NFC: $.062$, $p = .337$).

Analyzing Strategies and Actions

Since the tailorshop scenario is a dynamic simulation, participants experienced different situations depending on their ini-

tial decisions, making it hard to be captured and predicted by the limited causal map information. Therefore, we will now turn to the analysis of actions and strategies that participants used with respect to the resulting performance.

General Properties

Overall, 31.25% of the participants ended in debt, while only 10 participants (20.83%) had profitable tailorshops. For the following analyses, we focus on the difference between the profitable and all unprofitable tailorshops, not covering the differences to the subgroup of unprofitable tailorshops that ended up in debt specifically. First, the two groups are compared in terms of their expense, investment and income strategies. A breakdown thereof is shown in Figure 3. While the difference for the income is mostly due to the necessity of taking a loan or using up the savings, it becomes apparent that the profitable group very rarely relied on their savings over the course of the simulation. When considering the expenses and investments, the only major difference appears to be the investment in machines, which takes up a substantially larger proportion of the investments for the unprofitable group, and was invested in additional sales outlets instead by the profitable group. Overall, investment and expenses are rather similar, hinting at a problem of finding the right point in time: While comparable over the course of the run, the profitable group seems to make better decisions from the beginning (as indicated by the low proportion of used savings).

To gain a deeper understanding of the mechanisms causing the differences, we investigated the behavior of both groups on the level of actions and respective effects on the derived variables. Additionally, we included the exploration phase into the investigation, in order to see if participants with a better performance used the exploration phase to learn a strategy or had a better approach right from the beginning. Figure 4 shows the actions performed by both groups in each month as well as the resulting changes to the observable derived variables. From this, several findings are noteworthy: First, both groups had a negative outcome in the exploration phase, but used the phase by performing more drastic changes compared to the test phase, which was possibly the cause for the worse overall performance. Second, the first month showed by far the biggest changes and seemed to contain all the initial investments and adjustments that were planned to set the tone for the remaining months. Especially in the test phase, no substantial adjustments to the extend of the first month are made to any variable afterwards (besides selling everything right at the end to boost the final results). Third, the actions performed in the first month resembled the behavior already present in the exploration phase, with minor adjustments. This is corroborated by a significant strong correlation between the performance in the exploration phase with the performance in the test phase (One-sided Spearman's rank correlation: $r = .879$, $p < .001$).

Overall, a few clear but subtle differences between both groups emerged: For one, both groups switch to the machines with more capacity, but the profitable group sells the

old machines more decisive. For another, the profitable group seemed to avoid running in a supply shortage by investing more in raw material, machine maintenance, new machines and workers as well as salary compared to the unprofitable group. Finally, the unprofitable group starts with expanding outlets right away, which the profitable group is more hesitant to do. After the first month, the actions reflect the general situation: While the profitable group performs minor adjustments to advertisement and shirt price, the unprofitable group is forced to make cuts. While this is important for investigating the participants' ability to perform small-scale adjustments, it is mostly a product by the decisions that were made in the first month.

Predicting Performance

To predict the performance based on the actions, we rely on the same methods as in the causal map analysis. Again, we use the SVR, this time using the actions performed in the first month as inputs. The results (see Table 3) show that the SVR is now able to outperform ($MAE = 0.255$, $RMSE = 0.328$) the baseline models ($MAE = 0.295$ for the median baseline and $RMSE = 0.378$ for the mean baseline, respectively). Furthermore, it now achieves a positive coefficient of determination ($R^2 = 0.247$), indicating that, even for a simple general model, the first month provides easily accessible information.

Similarly to the correlation between causal map similarity and performance, we developed metrics aiming to correlate well with performance based on the actions. We used two simple heuristic strategies as metrics:

1. *Upgrade machines* (buy better machines, hire the respective workers, and sell the old machines), was calculated as follows:
   $strategy1 = sign(\Delta M100 + \Delta W100) * sign(-\Delta M50)$,
   where *sign* is the signum function and $\Delta M100$, $\Delta W100$ and $\Delta M50$ are the changes of the number of machines and workers.

2. *Avoid production loss* (buy raw material and invest in repair/maintenance), calculated as follows:
   $strategy2 = Material + Repair$

Both of the metrics correlated significantly with performance (Spearman's rank correlation for S1: $r = .310$, $p = .016$; for S2: $r = .656$, $p < .001$), indicating that a few actions in the first month are already good predictors for the performance. The present results suggest another explanation for the limited success of using causal maps: Due to the task's reliance on initial actions, many assumptions about variable interplay become irrelevant, whereas later decisions hard to predict due to the self-reinforcing and complex nature of the scenario.

Discussion

In the present article, we assessed three issues: First, we assessed the importance of the knowledge and assumptions that
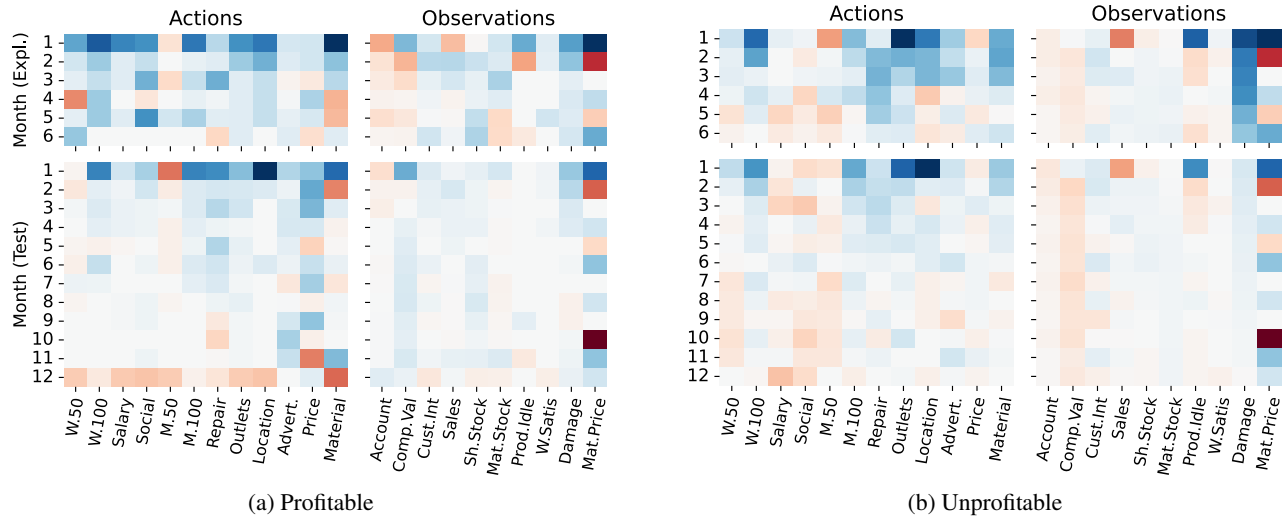
Figure 4: Average actions (changes to controllable variables) and the resulting observed variables for profitable/unprofitable tailorshops during the exploration and test phase. Darker shades of blue/red denote higher increments/decrements, respectively.

participants have before interacting with the tailorshop simulation are for the performance and if a revision of the assumptions is observable. For that, we obtained a causal map representing the relationships between the variables of the tailorshop before and after participants were interacting with the simulation. Thereby, the causal maps showed no significant signs that the initial assumptions were updated and could, despite correlating moderately, not be used as predictors of tailorshop performance when modeled using support vector regression. Similarly, the results of the Cognitive Reflection Task and the participants' Need for Cognition had no significant influence on the model. Finally, we found that the performance during the exploration phase was strongly correlated with the performance in the test phase, which further supports that no substantial changes to the assumptions were made. While it was expected that a real-world inspired scenario would be substantially impacted by real-world knowledge, part of the results could be explained by a limitation of the causal maps: The restriction to only represent positive or negative dependencies is too coarse to describe the dependencies that participants actually expect, introducing noise due to ambiguity and the lack of expressiveness.

Second, the actions selected by the participants were investigated. The analysis showed that the first month was by far the most dominant month, setting the tone for the whole run. This is likely to cause most other factors to become irrelevant, especially since the scenario itself is highly dynamic. Participants that made less fortunate actions in the first month rarely recovered, which in turn can likely alter their strategies. Although the first month is often excluded (Danner et al., 2011; Greiff et al., 2015), which is a reasonable means if the focus lies on the micromanagement feedback-loop during the other months, we argue that this is not ideal for cognitive modeling of complex problem solving. On the one hand, participants were instructed that the scenario has a time limit

of 12 months, where only the company value at the end mattered. This implies that each intermediate steps on its own does not necessarily reflect the actual thought processes. The fact that 25% of participants considered selling everything at the end to boost the final result further corroborates that they had, in fact, an overarching strategy. Since excluding the most impactful month from the performance evaluation strips the tailorshop almost entirely of its investment phase, in which planning and the strategies of participants arguably matter the most. Even when excluded, the first month will still alter the whole scenario and thereby the behavior of the participants, making it near impossible to predict. When predicting based on the initial actions, the support vector regression performed substantially better and outperformed the baselines. Furthermore, we were able to formulate simple strategies based on the first month alone that can serve as highly correlating predictors for success in the tailorshop. For predictive modeling endeavours, this leaves the tailorshop scenario in a tricky state, since most of the planning and adaption processes will be hidden by the dominant initial decisions, which can set the tone for the complete run in a complex non-linear scenario.

In conclusion, while we deem the use of complex problem solving in cognitive modeling important to extend its boundaries further into the area of real-world scenarios, we argue that the tailorshop gets trapped in its complexity, which makes it prone for snowball effects based on early actions. To this end, our results align with the critique by Greiff and Funke (2009) on "one-item-testing". Especially for cognitive modeling, it is essential to rely on a complex tasks that is either easily repeatable (i.e., by having multiple items), or is less self-reinforcing, so that the actions performed by participants across all steps of the tasks have a similar impact. In the end, overarching strategies have to be observed at the same time as small step-to-step adjustments — since both are essential components of real-world complex problem solving.

Acknowledgements

References

Awad, M., & Khanna, R. (2015). Support vector regression. In *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 67–80). Berkeley, CA: Apress. doi: 10.1007/978-1-4302-5990-9_4

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2015). Deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition NFC-K [German short scale for measuring the construct Need for Cognition NFC-K]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*.

Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., & Funke, J. (2011). Measuring performance in dynamic decision making. *Journal of individual differences*, *32*, 225-233. doi: 10.1027/1614-0001/a000055

Dörner, D., & Wearing, A. J. (1995). Complex problem solving: Toward a (computersimulated) theory. In *Complex problem solving* (pp. 65–99). Psychology Press.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processings. *Human factors*, *4*(2), 59–74.

Funke, J. (1988). Using simulation to study complex problem solving: A review of studies in the FRG. *Simulation & Games*, *19*(3), 277–303.

Funke, J. (2014). Problem solving: what are the important questions? In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 493–498).

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in human behavior*, *21*(2), 273–286.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving: The microdyn approach. Office for Official Publications of the European Communities.

Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence*, *50*, 100-113. doi: https://doi.org/10.1016/j.intell.2015.02.007

Pitz, G. F., & Sachs, N. J. (1984). Judgment and decision: Theory and application. *Annual review of psychology*, *35*(1), 139–164.

Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [The relation between test intelligence and problem solving success]. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*, *189*(1), 79-100.

Putz-Osterloh, W. (1983). Über Determinanten komplexer Problemlöseleistungen und Möglichkeiten zu ihrer Erfassung [On factors for complex problem solving and possibilities of their diagnosis]. *Sprache & Kognition*, *2*, 100–116.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147-168. doi: 10.1080/13546783.2013.844729

Dissecting the Drivers of Change Points in Individual Learning: An Analysis with Real-World Data

Michael Collins (michael.collins.74.ctr@us.af.mil) NRC Postdoc at AFRL Dayton, OH, USA

Florian Sense (florian.sense@infinitetactics.com) InfiniteTactics, LLC Dayton, OH, USA

Michael Krusmark (michael.krusmark.ctr@us.af.mil) CAE, Inc. Wright Patterson Air Force Base, Ohio

Tiffany Jastrzembski (tiffany.jastrzembski@us.af.mil) Air Force Research Laboratory Dayton, OH, USA

Abstract

Many different theories of learning have been developed to account for human performance over time, often accounting for performance at an aggregate level. Understanding performance at an individual level is often more difficult because of multiple different factors (e.g., noise, strategy selection, or change in memory representation), which are often not accounted for in simple learning theories. One approach used to explain the sudden changes in performance that are often observed at the individual level is to integrate change detection algorithms with psychological models. This research has shown that performance at the individual level can be understood not by a single continuous process but instead by segmented portions of multiple processes. Previous research has posited different explanations as to what features drive the inferences of change points. However, no paper has yet compared different explanations' ability to explain the variance in inferred change points. In this paper, we use a simple model of learning to account for performance in a real-world data set with individuals performing multiple different games that tap into different task attributes (i.e., memory, attention, problem-solving) on the website Luminosity. We then conduct a statistical analysis to determine what drives change points in the dataset. The results here allow for better clarification as to what features are driving the inferences of change points at the individual level.

Keywords: Learning, Change detection, Real-World data, Cognitive Models

Introduction

One primary interest in psychology has been understanding how individuals learn and retain information over time across a range of tasks. Research on learning has spanned various levels of complexity from simple paired associates tasks (Newell & Rosenbloom, 2013) to complex tasks (Gray & Lindstedt, 2017), illuminating how individuals learn and acquire information about general domains (e.g., algebra) and develop high levels of expertise (e.g., chess), (Newell & Simon, 1972). To explain human learning across these different domains, multiple different models have been developed according to different features of memory (e.g., decay, spacing effect, Kumar, Benjamin, Heathcote, & Steyvers, 2022), learning mechanisms (procedural vs declarative memory), and strategies (instance-based learning, Gonzalez, Dutt, & Lebiere, 2013).

These different learning models often focus on average performance across individuals. Although, the focus on accounting for average human performance certainly has a place in psychology, it can lead to the development of models that do not generalize to individual-level performance (Estes &

Todd Maddox, 2005; Heathcote, Brown, & Mewhort, 2000). The difficulty in generalizing models developed to fit average to individual performance stems from the fact that individual performance often does not follow smooth, continuous performance curves that best account for average performance, but instead contain periods of improvement followed by sudden increases or decreases in performance. The sources of variability in individual performance can be thought of as stemming from different features: (1) individual differences, such as learning rates (Lee, Gluck, & Walsh, 2019; Heathcote et al., 2000), (2) problem-solving strategy (Gray & Lindstedt, 2017), or (3) learning mechanisms (Smith & Minda, 1998).

To account for the variability in performance at the individual level, cognitive models have been paired with change detection algorithms, to estimate change points in an individual's performance to account for shifts in the underlying process (e.g., change in model - Lee et al., 2019; Tenison & Anderson, 2016 or model parameters, Tenison & Anderson, 2016; Collins, Sense, Krusmark, & Myers, 2023). Change detection algorithms are statistical approaches that attempt to identify homogeneous segments within time series data (Serre, Chételat, & Lodi, 2020). Previous research has shown that the variability in individual performance is not best explained by simply noise but systematic variation in learning trends (Gray & Lindstedt, 2017; Lee et al., 2019; Tenison & Anderson, 2016; Collins et al., 2023). Though this prior research has been able to explain individual performance at a fine level of granularity, different explanations as to what features of an experiment give rise to these estimated changes in performance have been given. In this paper, we review three different explanations as to why change points are inferred and evaluate their ability to explain the estimates of change points in a real-world dataset.

Explanations for Change Points

The research that has explored integrating change detection algorithms with psychological models has provided different explanations as to why changes are inferred while learning (Gray & Lindstedt, 2017; Tenison & Anderson, 2016; Lee et al., 2019). One explanation posited by Gray and Lindstedt (2017) is that sudden shifts in performance are the result of changes in the individual's strategy on a particular task. Changes occur during skill acquisition when individuals explore and refine problem-specific strategies to solve a

task which may allow the individual to improve their performance compared to a previously used strategy. Positing strategy exploration allows for models of learning to account for performance profiles that include consistent improvements as well as periodic increases or decreases in performance, which can be understood as an individual trying a new strategy. Given Gray and Lindstedt (2017)'s explanation of why change points are inferred two predictions are made. First, change points are dependent on the task. Second, change points are equally likely to occur at any time while performing a task depending on the experience and motivation of the individual while performing the task.

A second explanation for why changes occur posits that sudden changes in performance are due to memory consolidation. Tenison and Anderson (2016) argue that individuals progress through three stages when learning to solve a particular problem: First the solution to a task is represented using purely declarative memory; next, the declarative information is consolidated into a mix of declarative and procedural memory representations; and finally, the task information is compiled into procedural memory. Each of these phases of learning is assumed to have a unique learning curve (Tenison & Anderson, 2016; Kim, Ritter, & Koubek, 2013). Given Tenison and Anderson (2016)'s explanation of change points, two predictions are made. First, inferred change points are dependent on a problem's content, which must be able to be consolidated from declarative to procedural memory. Second, change points are more likely to be inferred early in performance and then decrease over time as the individual moves through the three stages of learning.

Finally, Lee et al. (2019) posit that sudden changes in performance occur when individuals explore different strategies due to inherent motivation or shifts in the payoff structure of the environment. However, in contrast to Gray and Lindstedt (2017), Lee et al. (2019) assume that the number of inferred change points is a function of the individual—not specific task or problem content. Based on this, two predictions are made. First, the number of inferred change points should vary between individuals and not be dependent on the task or problem content. Second (like Gray & Lindstedt, 2017), change points can occur at any point in time due to strategy exploration.

The Current Work

In summary, the work reviewed above proposes that change points occur for different reasons: the specific problem being learned (Gray & Lindstedt, 2017), problem-content (Tenison & Anderson, 2016), and participant (Lee et al., 2019). Previously, each of these explanations has been applied to different experimental situations: Space Fortress, novel math problems, and decision-making tasks, respectively. However, no attempt has been made to compare these different explanations against each other on a single dataset. In this paper, we compare the ability of the three different explanations to best account for inferred change points applied to a real-world dataset of individuals completing different games on Lumi-

nosity (Steyvers & Schafer, 2020). Specifically, our analyses will zoom in on the following distinguishing points:

- Are the number of change points inferred across all game plays on Luminosity best *explained* and *predicted* by either the subject, game attribute or the specific game?

- How does the probability of a change point occurring change over game plays?

The remainder of this paper is structured as follows. First, we give an overview of the Luminosity dataset (Steyvers & Schafer, 2020) used in this paper. Second, we outline the learning model and procedure used to infer change points in the individuals' performance on Luminosity. Third, a series of analyses based on the inferred change points is presented to shed light on the above question. Finally, we explore the implications of our results and outline areas of future research.

Method

Luminosity Dataset

Luminosity is an online learning platform where individuals can play games to improve various cognitive abilities. The dataset collected by (Steyvers & Schafer, 2020) consisted of a record of up to the first 60 game plays of 36,297 individuals playing 84 different games. All of the recorded data from Luminosity was performed online as opposed to on a mobile phone. A latent factor analysis was used to categorize games as having one of six primary attributes: attention, problem-solving, memory, flexibility, speed, or math (Steyvers & Schafer, 2020).

Participants The full Luminosity data (Steyvers & Schafer, 2020) was composed of 36,297 individuals (Male = 39%, Female = 50%, Unidentified = 11%) ranging from 18-91 years old with a range of education. For this paper, 214 participants were randomly sampled from the full dataset for our analysis. A fairly small subset of the full dataset was used in this because of the computational time required to run the change detection algorithm.

Performance Measures On Luminosity, participants received a score based on their (1) accuracy, (2) speed, and (3) bonus points after each time they played a game. The Luminosity dataset contained the participants' raw performance scores as well as normalized performance scores[1]. For the analysis conducted in this paper, the normalized performance scores were used so that performance across different games could be compared.

Learning Model

Various learning models have been developed and compared both for the Luminosity data (Kumar et al., 2022) and in conjunction with change detection algorithms (Gray & Lindstedt,

---

[1]See Steyvers and Schafer (2020) for details on how performance scores were normalized.

2017; Tenison & Anderson, 2016; Lee et al., 2019). Here, we use a well-established exponential learning model:

$$\text{Performance} = A - U \times e^{-N \times \alpha} \qquad (1)$$

The model's fixed parameter (N) represents the number of exposures on a particular task. While the model's free parameters control the maximum performance value (A), the performance intercept (U), and the rate of learning (α).

Model Fitting

For this paper, the above learning model was paired with a change detection algorithm (Serre et al., 2020) fit individually to each game played by each participant using a genetic algorithm. Genetic algorithms are a type of optimization algorithm that mimic natural selection. They work by specifying a 'population' of potential parameter values (i.e., possible change points) and then determining the 'fitness' of each population (e.g., BIC, RMSD, r, likelihood, etc.). After the fitness of a set of parameters (i.e. population) has been determined, a new set of potential parameters is generated according to mutation, cross-over, and fitness values based on the previous population. The process of recursively modifying and specifying a set of parameter values is repeated until the algorithm converges on a solution. One key parameter that needs to be set is the maximum number of change points that can be inferred. Here, we used the total number of game plays completed by the participant on a particular Luminosity game.

This approach was combined with the cognitive model in the following way: First, the genetic algorithm proposed a set of potential change points that separated the participant's performance into different segments. Next, the learning model's three free parameters (A, U, α) and a parameter for the standard deviation of a normal distribution (SD) were fit separately to each segment via maximum likelihood. Then, a fitness value for the combination of proposed change points *and* the learning model's parameters was determined. Convergence is declared if the fitness is not improved for ten consecutive iterations. We chose the BIC as our measure of fitness because it takes into account (1) the number of free parameters (i.e., number of change points and the learning model's free parameters), (2) the likelihood of the learning model's fit to each segment, and (3) the total number of game plays. Each additional change point that is added thus adds five additional parameters (the change point, A, U, α, and SD, which is used to compute the likelihood). The BIC ensures that the increased complexity of adding more change points is warranted by a proportional increase in 'fitness.'

Results

Overall Fit and Inferred Change Points

First, we review the model's fit to the participants' performance across all Luminosity games and the number of inferred change points inferred across all games. The model's fit was assessed using correlation (r) and root mean squared error ($RMSD$). Overall, the learning model paired with the change detection algorithm was able to fit the participants' performance very well ($r = .95$, $RMSD = 8.56$) (Figure 1). An examination of the average normalized performance over time reveals a standard average performance curve with low initial performance that increased over time reaching a performance plateau (Figure 1).
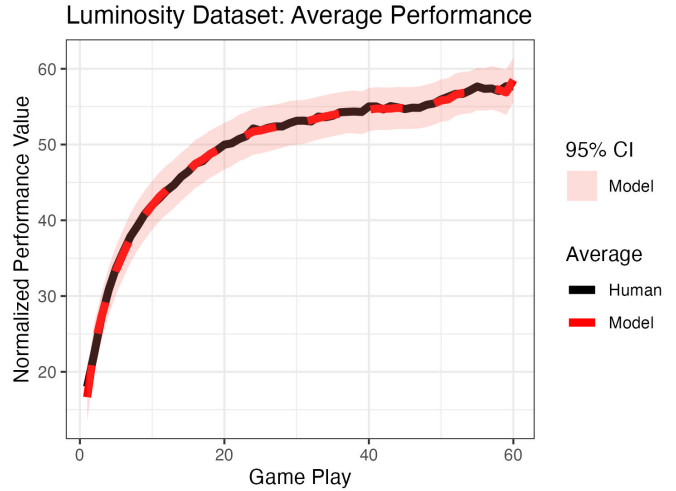
Figure 1: Average human performance (black line) across all game plays and the model's fit (dashed red line and ribbon +/- 95% CI).

Next, we examined the total number of inferred change points across all of the games played on Luminosity. As can be seen in Figure 2 at least one change point was inferred ($M = 2.84$, $SD = 2.12$) for the majority of the participants (90.9%). Furthermore, the distribution of the number of change points is highly skewed, showing that a large number of change points was uncommon.
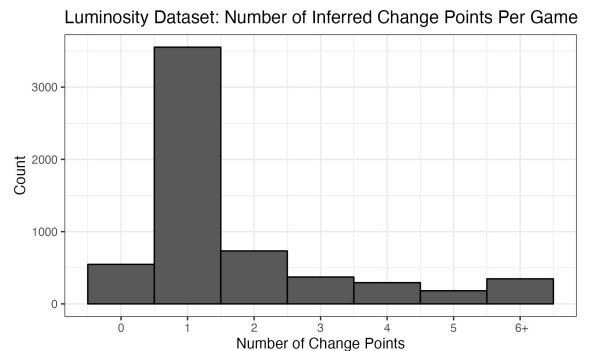
Figure 2: Histogram of the number of inferred change points per game across all participants.

Sources of Variance in Change Points

Addressing our first key question directly, we want to determine which features explain variance in the occurrence of

change points. To this end, we used the sequence of game plays associated with each user playing the various games and fit a logistic mixed-effects regression model. The dependent variable was the binary result of the change detection algorithm: Was a change point inferred for this game play? As a first step, we fit an intercept-only model that included random intercepts for each of three variables that correspond to the possible sources of variance outlined in the introduction: the user, the game, and the primary attribute (e.g., 'memory') associated with each game. We compared the intercept-only model with random effects for all three variables to the intercept-only model with no random effects as well as all possible combinations of said effects. Both the *AIC* and *BIC* favor the model that includes both user- and game-specific random intercepts. This model narrowly outperforms the model that also includes random effects for the primary attribute (*BIC* 77,959 vs. 77,970 and *AIC* 77,929 vs. 77,930). By looking at the adjusted intra-class correlation (*ICC*) values (Lüdecke, Ben-Shachar, Patil, Waggoner, & Makowski, 2021) for the intercept-only models, we learn how much of the variance in whether a change point is inferred is attributable to the random effects. *ICC* values are 0.218 for the user-game model and 0.219 for the model with all three random effects, confirming that there is a tiny advantage of the full model[2].

Given prior research, however, the model likely needs to be extended to account for the likelihood of change points as a function of practice (i.e., game plays). Speaking directly to our second key question: Change points were more likely to occur during earlier game plays (see black dots in Figure 3). The model fit shown in Figure 3 is an extension of the intercept-only model above to which we added fixed effects for *game play number* (scaled but shown on the original scale for clarity). Further model comparison suggests that adding a quadratic term for *game play number* further improves model fit (*AIC*: linear = 77,329 vs quadratic = 77,057; *BIC*: linear = 77,369 vs quadratic = 77,108). Adding a third-degree polynomial was not warranted. The depicted model is the end result of a further model comparison step in which we tested the different random effects structures in the presence of the quadratic term of *game play number*. The comparison confirmed the results based on the intercept-only model: Both user- and game-specific random effects explain significant amounts of variance in change point occurrence but adding attribute-based effects only has a tiny benefit that is not warranted given the added complexity.

Taken together, this set of model comparisons speaks to previous explanations given for change points. We see that change points do indeed occur more frequently early on, which would be in line with the assertion by Tenison and Anderson (2016). However, in these data, an uptick in the probability of a change points is apparent in the final game plays,
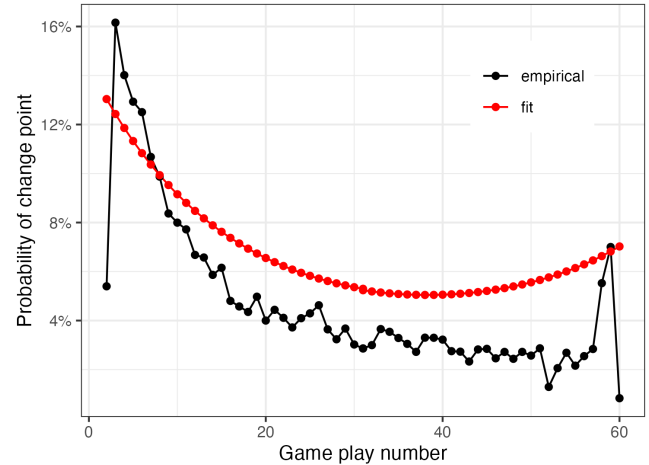
Figure 3: The ratio of inferred change points per game play (black line) and the linear mixed model's fit (red line).

which is why the quadratic term in the model improved the fit markedly. This in turn goes against the purely negative relationship between change points occurrences and time, as proposed by Gray and Lindstedt (2017) and Lee et al. (2019). Furthermore, this analysis also provides some insight into the dominant sources of variability in the occurrence of change points. There is no strong evidence that the primary attribute of the task (e.g., memory vs attention) provides significant explanatory power while both user- and game-based sources of variance do.

Predicting the occurrence of change points

For the final analysis, we will adopt a predictive rather than explanatory approach (Shmueli, 2010). The goal of this analysis is to shed more light on the sources of variance that are leveraged when the goal is to predict whether a change point will be inferred on the next trial. This is not unlike the mixed-effects regression approach in the previous section. Here, however, we will leverage a gradient-boosted decision tree ensemble that can learn and leverage non-linear relationships between the predictors without requiring us to specify them in advance. The feature importances extracted from the trained model are our primary interest.

Specifically, we trained an XGBoost classifier using log-loss (max depth = 6; $\eta = 0.2$; rounds = 100; using the implementation from Chen et al., 2024). As above, the outcome measure was whether a change point is inferred for a given game play number. As predictors, we one-hot encoded the users, games, and primary attributes. Additional predictors were the *game play number*, the lagged normalized performance (we included lags one through three as predictors), as well as sliding window mean performances (with window sizes three and nine; all based on lag-1 performance to avoid data leakage).

We trained the model on 80% of the data (log loss = 0.177) and evaluated it on the 20% hold-out set (log loss = 0.192)

---

[2]For reference, the *ICC* values for the models that only include a single random effect each are: user = 0.085; game = 0.112; attribute = 0.032.

to verify that it generalized. From the trained model, we extracted the gain for each feature as a metric of feature importance. Because we were primarily interested in groups of features (e.g., how important are users as a grouping relative to games as a grouping), we used model-based feature aggregation: First, we scaled the gain values so they sum to 1; then we summed all scaled gain values for each grouping to get an aggregate importance value. For example, there are 84 games. Their labels were turned into 84 one-hot encoded predictive features, each of which was associated with a gain value (i.e., feature importance). Summing the scaled gain values for all 84 games gives a value of 0.20, which is the proportion of overall importance attributable to the feature category 'game.' The proportions of model-based feature aggregations are shown in Figure 4. The exception in this graph is *game play*, which is only a single feature[3]. We can see from the graph that performance-based features are most important, the top three of which are the lag-1, sliding window mean with size 9, and lag-2 (proportions: 0.129, 0.078, and 0.075, respectively).

Broadly speaking, this predictive modeling approach confirms the explanatory analyses above. Both user- and game-based features are more important than the primary attribute—which contributes to the predictive power but not much[4]. Game play is the single most important predictor, which is in line with the results in Figure 3. Notably, the performance in preceding trials—which is what the lagged values represent—is the most important category of predictors. Naturally, the change detection algorithm that determines the change points that are predicted here depends on the preceding performance values. Hence, it is not too surprising that the pattern of scores just before the change point contain predictive information. One downside of using a machine learning model like XGBoost that can learn arbitrary mappings between the predictors is that the final model usually contains non-linear, higher-order interactions between features that are difficult to disentangle.

Discussion

Previous research attempting to understand learning at an individual level has explored integrating change detection algorithms with psychological models of learning (Gray & Lindstedt, 2017; Lee et al., 2019; Tenison & Anderson, 2016). Pairing change detection algorithms with models of learning has allowed psychological models to better account for the variability found in individual performance. However, despite the improved ability to account for individual performance, varying explanations as to what features are most associated with inferred change points (participant, problem-type, or problem) and when they will be inferred (i.e., early

---

[3]We did not include specific transformations or polynomials as in the previous analysis because the tree ensemble is well-equipped to learn any (potentially non-linear) patterns directly from the data.

[4]The two most important attributes are 'problem solving' and 'math' with proportions 0.021 and 0.010, respectively. All other attributes have values < 0.01.
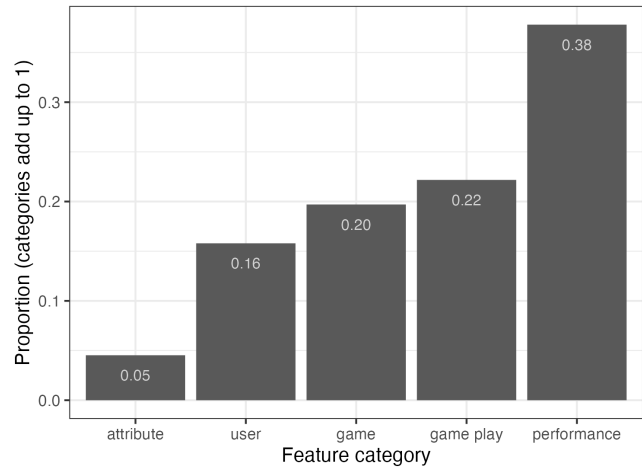
Figure 4: Model-based feature aggregation extracted from the trained XGBoost model.

on in performance or uniformly over time) have been given. For this reason, we compared these different explanations to see which features could best explain the variance in inferred change points in a real-world dataset. Our analysis focused on what features of the data would best explain and predict the inferences of change points across individuals' performance.

Our results revealed that the variation in inferred change points could be understood by previously proposed explanations (Gray & Lindstedt, 2017; Tenison & Anderson, 2016; Lee et al., 2019) to varying degrees. We found that the most relevant features for explaining the variability in change points were user and game, in line with the explanations given by Gray and Lindstedt (2017) and Lee et al. (2019). Little evidence was found that problem-type accounted for meaningful variation across the inferred Luminosity dataset, which went against the prediction made by Tenison and Anderson (2016). These findings were corroborated by the predictive modeling approach, which also identified user and game as being markedly more important than problem-type (i.e., primary attribute).

Furthermore, we examined how the probability of change points changed over game plays. We found that there was a non-linear relationship between change points and game plays, with most change points being inferred during early game plays and a slight uptick during later game plays. The initial high number of change points during initial game plays is consistent with the explanation given by Tenison and Anderson (2016). However, the slight but significant increase in inferred change points during later game play is consistent with the explanation given by Lee et al. (2019) and Gray and Lindstedt (2017).

Finally, using a gradient-boosted decision tree ensemble, we found that the change points inferred by that change detection algorithm could be predicted, using a variety of different features, such as prior performance, game play, game, and user. This result further supports the notion that the inferred

change points are systematically identifiable and predictable based on past performance.

Limitations and Future Research

Several limitations and lines of future research should be acknowledged. First, the inference of change points will always depend on the model and change point algorithm selected. For this paper, a simple learning model was chosen to fit the participant's performance. However, more complex cognitive models of learning (e.g., PPE, Collins et al., 2023) might be able to better account for performance over longer periods leading to the inference of fewer inferred change points. Second, this paper focused solely on learning over time based on the number of exposures to a particular game. However, other cognitive mechanisms such as memory decay also play a role in learning over time and have shown to be relevant for accounting for performance over time on Luminosity (Kumar et al., 2022; Collins et al., 2023). Future research should explore the robustness of these findings when using different change detection algorithms and more complex models of learning to see if the findings reported in this paper can still explain the inferred change points across individuals' performance

Conclusion

Incorporating change detection algorithms with psychological models allows for models to better account for large amounts of variability in human performance. The benefit of this approach is that more nuanced theories can be evaluated at an individual level to better understand human learning. However, incorporating change detection algorithms with psychological models also increases the complexity of models. Our results found that the explanations previously given can explain the variability in inferred change points in a real-world dataset to various degrees. These findings both support the conclusions of previous work and also provide an opportunity for model development to better explain human learning at an individual level.

Acknowledgments

References

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... Yuan, J. (2024). xgboost: Extreme gradient boosting [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=xgboost (R package version 1.7.7.1)

Collins, G. M., Sense, F., Krusmark, M., & Myers, T. (2023). Modeling change points and performance variability in large-scale naturalistic data. MathPsych/ICCM.

Estes, W. K., & Todd Maddox, W. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic bulletin & review*, *12*(3), 403–408.

Gonzalez, C., Dutt, V., & Lebiere, C. (2013). Validating instance-based learning mechanisms outside of act-r. *Journal of Computational Science*, *4*(4), 262–268.

Gray, W. D., & Lindstedt, J. K. (2017). Plateaus, dips, and leaps: Where to look for inventions and discoveries during skilled performance. *Cognitive science*, *41*(7), 1838–1870.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, *7*(2), 185–207.

Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science*, *14*(1), 22–37.

Kumar, A., Benjamin, A. S., Heathcote, A., & Steyvers, M. (2022). Comparing models of learning and relearning in large-scale cognitive training data sets. *npj Science of Learning*, *7*(1), 24.

Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*, *6*(4), 335.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. doi: 10.21105/joss.03139

Newell, A., & Rosenbloom, P. S. (2013). Mechanisms of skill acquisition and the law of practice. In *Cognitive skills and their acquisition* (pp. 1–55). Psychology Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Serre, A., Chételat, D., & Lodi, A. (2020). Change point detection by cross-entropy maximization. *arXiv preprint arXiv:2009.01358*.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, memory, and cognition*, *24*(6), 1411.

Steyvers, M., & Schafer, R. J. (2020). Inferring latent learning factors in large-scale cognitive training data. *Nature Human Behaviour*, *4*(11), 1145–1155.

Tenison, C., & Anderson, J. R. (2016). Modeling the distinct phases of skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 749.

The Computational Mechanisms of Detached Mindfulness

Brendan Conway-Smith (brendan.conwaysmith@carleton.ca),
Robert L. West (robert.west@carleton.ca)
Department of Cognitive Science, Carleton University, Ottawa, ON K1S 5B6 Canada

Abstract

This paper investigates the computational mechanisms underlying a type of metacognitive monitoring known as detached mindfulness, a particularly effective therapeutic technique within cognitive psychology. While research strongly supports the capacity of detached mindfulness to reduce depression and anxiety, its cognitive and computational underpinnings remain largely unexplained. We employ a computational model of metacognitive skill to articulate the mechanisms through which a detached perception of affect reduces emotional reactivity.

Keywords: metacognition; mindfulness; affect; emotion; proceduralization; ACT-R; Common Model

Introduction

The attempt to build a Unified Cognitive Architecture (Newell, 1994) that can replicate human-like intelligence must necessarily account for the routine interplay between affect and metacognitive processes. Historically, cognitive modeling research has focused predominantly on knowledge-based processing such as reasoning, vision, and AI problem-solving, with little or no computational account of the critical role of emotion and metacognition.

This need for increased computational understanding is underscored by the fact that perseverative patterns of negative emotion, such as depression and anxiety, are the largest causes of cognitive disability worldwide (World Health Organization, 2022). Consequently, there has been a global push to develop metacognitive techniques that allow individuals to engage with their emotions adaptively. A particularly effective metacognitive technique is referred to as 'detached mindfulness' (Wells, 2005). This technique focuses on developing one's perception of the momentary changes of affective states, shown to significantly reduce feelings of distress, emotional reactivity, and to improve overall cognitive functioning (Hammersmark et al., 2024).

While decades of clinical research strongly supports the effectiveness of metacognitive strategies and detached mindfulness in particular, their underlying cognitive and computational mechanisms remain largely unexplained. This paper will investigate the cognitive and computational constituents that underpin detached mindfulness and its therapeutic benefits. Specifically, we will discuss the metacognitive mechanism by which the perception of affective fluctuations deactivates emotional reactivity.

For this purpose, we will employ the Common Model of Cognition (CMC), originally the 'Standard Model' (Laird, Lebiere, & Rosenbloom, 2017), which provides a unified framework for investigating the fundamental elements of cognitive and metacognitive phenomena. By utilizing the Common Model, and specifically ACT-R (Anderson & Lebiere, 1998) in this investigation, we intend to address important questions largely unexplored in cognitive models: How does metacognitive training in detached mindfulness reduce perseverative styles of negative emotions? By what computational mechanism does perceiving the momentary changes in affect disengage emotional reactivity such as meta-emotions?

First, we will overview the relevant literature on metacognition and mindfulness techniques. Second, we outline the computational mechanisms involved in a model of metacognitive skill learning. Third, we apply this model of metacognitive skill learning to detached mindfulness to clarify its underling components and the precise mechanism by which it reduces emotional reactivity as reported in the literature.

Metacognition

We propose that an active mechanism of detached mindfulness fundamentally relies on a form of automatized metacognition. The common conception of metacognition refers to the monitoring and control of cognitive processes (Flavell 1979; Fleming, Dolan, & Frith, 2012). *Metacognitive control* refers to the active regulation of cognitive processes or states to either activate or inhibit them (Proust, 2013; Wells, 2019). The regulation of one's own cognitive processes can involve various processes such as attention, emotion, planning, reasoning, and memory (Efklides, Schwartz, & Brown, 2017; Pearman et al., 2020). *Metacognitive monitoring* refers to the capacity to recognize and identify cognitive states. It involves the perception of internal mental states such as thoughts and feelings in order to regulate those states or direct behavior.

Studies demonstrate that metacognitive monitoring can be developed and improved through training (Baird, Mrazek, Phillips, & Schooler, 2014). For instance, attentional processes can be developed and enhanced through the repeated practice of attention-based tasks (Posner et al., 2015). Metacognitive training such as mindfulness techniques is integral to both Cognitive Behavior Therapy (CBT; Dobson, 2013) and

Metacognitive Therapy (MCT; Normann & Morina, 2018) and facilitates improved control over maladaptive thoughts and emotions (Wells, 2011, 2019; Hagen et al., 2017). The benefits of mindfulness training rely partly on its enhancement of metacognitive sensitivity, which is the extent to which one is able to perceive their own mental processes or states, including thoughts, feelings, and emotions (Fleming & Lau, 2014). Improved metacognitive sensitivity has the effect of lowering one's metacognitive threshold — the minimal level of a stimulus required for a person to be aware of some mental state and make a judgment about it (Charles, Chardin, & Haggard, 2020; Pauen & Haynes, 2021). The metacognitive threshold can also be lowered by way of attentional training, such as detached mindfulness and meditation, which allows one to perceive a weaker signal strength from internal cognitive states (Fox et al., 2016). While this has been effectively modelled within ACT-R (Conway-Smith & West, 2023) it is not the main focus of this paper.

Metacognition as mindfulness

Scientific interest in mindfulness practice has become a target of interdisciplinary research and has grown exponentially over the past few decades (Van Dam et al., 2018). Metacognition and mindfulness are often used interchangeably within cognitive psychology (Holas & Jankowski, 2013). Mindfulness psychology contends that a significant degree of emotional distress and pathological symptoms are caused by the illusory perception of affective experience being more permanent than it actually is. This perceptual illusion has been explained as the result of poor metacognitive sensitivity that obscures the detection of affective fluctuation (Brown & Ryan, 2003; Grossman et al., 2010). To address this metacognitive deficiency, detached mindfulness has emerged as a uniquely effective therapeutic technique (Wells & Matthews, 1994; Hammersmark et al., 2024). This involves participants learning to observe moment-to-moment changes in mental states, including subtle emotional fluctuations, and allowing these states to occur without engaging with or reacting to them.

This non-reactive state of awareness is also referred to as 'equanimity'. In mindfulness therapies that do not promote equanimity, awareness alone is often insufficient to increase subjects' psychological well-being (Cardaciotto et al., 2008). Detached mindfulness is most closely aligned with Vipassana meditation (in the tradition of S.N. Goenka), an old and popular technique that largely focuses on cultivating equanimity i.e., perceptual sensitivity to variations in affect and physical sensation (Kakumanu et al., 2018). Regular practice of this technique has shown to improve executive functioning, response inhibition, and control over emotional reactions such as meta-emotions (Andreu et al., 2019).

Meta-emotion

Meta-emotions are emotions that automatically react to other emotions (Jäger & Banninger-Huber, 2015; Predatu, David & Maffei, 2020). For instance, a primary negative emotion (sadness) can cause a greater secondary negative emotion (despair) which may cause an even greater tertiary negative emotion (depression). Meta-emotions are instances of positive feedback, in which an emotional response to a primary emotion intensifies the overall emotional experience, leading to an amplified response. Meta-emotions occur as low-level reactive processes that are largely unconscious and involuntary, making them difficult to intervene in.

While therapeutic practices aim to control the resulting effects of meta-emotions such as anxiety and depression, techniques such as detached mindfulness and Vipassana aim to address the source, which is considered the false perception of affective permanence. To date, we lack a mechanistic understanding of precisely how detached mindfulness breaks through the illusion of affective permanence and disengages emotional reactivity. To clarify this mechanism, we will apply a model of metacognitive skill learning that articulates the components involved in this process and how they interact. Central to this explanation is a process referred to as proceduralization, a framework that is common among skill theories. We will first discuss the relevant components of metacognition and their expression in the cognitive architecture ACT-R. We will then explore how the components of proceduralization function to produce the therapeutic mechanism active in detached mindfulness.

Components of metacognition

There are at least two types of cognitive representations that can engage in metacognitive monitoring and control processes — declarative knowledge and procedural knowledge. Metacognitive knowledge, or meta-knowledge, is considered a form of declarative knowledge (Schraw & Moshman, 1995; McCormick, 2003; Wells, 2019). Meta-knowledge takes the form of an explicit metarepresentation that is propositionally formatted and refers to a cognitive property, e.g.: "I am focused" (Shea et al., 2014; Proust, 2013). Meta-knowledge can also take the form of a metacognitive instruction, which specifies a mental action to be performed (Wells, 2019). A metacognitive instruction, or meta-instruction, prescribes an action directed toward controlling some cognitive process, e.g.: "Focus on the current task." Metacognitive knowledge is considered to be distinct from metacognitive skill, as it does not automatically lead to the deployment of metacognitive processes (Veenman & Elshout, 1999).

The execution of metacognitive instructions is performed by way of procedural knowledge. Improvements in metacognition are said to involve the

refining of procedural knowledge that people use to monitor and control their own cognitive processes (Brown & DeLoache, 1978; Schraw & Moshman, 1995; Wells, 2019). The various realms of metacognitive skills can be understood as different domains of procedural knowledge (Veenman et al., 2005).

ACT-R

Various theories of metacognition have been modelled within the ACT-R cognitive architecture (Reitter, 2010; Anderson & Fincham, 2014). ACT-R instantiates decades of research on how human cognition functions computationally. Its mandate is to depict the components necessary for human intelligence, which include working memory, perception, action, declarative memory, and procedural memory. These modules have also been correlated with their associated brain regions (Borst et al., 2015).

The ACT-R cognitive architecture fundamentally distinguishes between declarative and procedural knowledge, which accords with the literature on skill acquisition in philosophy and psychology (Squire, 1992; Christensen, Sutton, & McIlwain, 2016). Declarative knowledge is formatted propositionally and structured within semantic networks. Procedural knowledge is commonly referred to by researchers as containing "procedural representations" (Anderson, 1982; Pavese, 2019). Within ACT-R, procedural representations are computationally specified as "production rules" which are a dominant form of representation within accounts of skill (Newell, 1994; Taatgen & Lee, 2003; Anderson et al., 2019). Neurologically, production rules are associated with the 50ms decision timing in the basal ganglia (Stocco, 2018). Production rules, or "productions", transform information and change the state of the system to complete a task or resolve a problem. A production rule is modeled after a computer program instruction in the form of a "condition-action" pairing. It specifies a condition that, when met, performs a prescribed action. A production is also thought of as an "if-then" rule. *If* the condition is satisfied, such as matching to working memory, *then* it fires an action (Figure 1).

Figure 1: Production rules are formatted as a condition-action pairing. IF the condition side matches to the cue in working memory, THEN it fires an action.

Affect have been modelled computationally within ACT-R as non-propositional representations in working memory, or "metadata" (West & Conway-Smith, 2019).

These types of affective information, encompassing both emotional states and noetic feelings, are regarded essentially as patterns within working memory that can be accessed by production rules.

Production rules match to and fire off the content in working memory. Should any stimuli or pattern appear in working memory, productions that match this pattern will arise from procedural memory and fire a prescribed action. In this way, cues in working memory can prompt procedural knowledge to act within various domains; motor, cognitive, and metacognitive. It is these specific cognitive units that are developed and refined during the process of proceduralization.

Proceduralization

The concept of proceduralization is often used within the skill acquisition literature to explain the cognitive mechanisms involved in task learning (Fitts & Posner, 1967; Dreyfus & Dreyfus, 1986; Kim & Ritter, 2015). It refers to the process by which a task becomes automated, allowing it to be performed more efficiently and accurately, with minimal conscious effort or attention. The process involves converting slow declarative knowledge into fast procedural knowledge which is then increasingly refined. Skill performance can be further improved by way of mechanisms such as time delayed learning, where faster productions are rewarded. Proceduralization plays a significant role in the cognitive processes involved in skill learning within domains such as motor skill, cognitive skill, and metacognitive skill (Fitts, 1964; Anderson, 1982).

Metacognitive proceduralization

Metacognitive proceduralization involves a mechanism by which human cognition can become more skillful at monitoring and controlling its own processes, such as attention, emotion, and metacognitive sensitivity (Conway-Smith, West, & Mylopoulos, 2023). Previous research has presented proceduralization as a mechanism that can lower the metacognitive threshold, allowing one to perceive increasingly weaker signals from mental states and more subtle changes in affect (Conway-Smith & West, 2023). It is hypothesized that proceduralization accomplishes this through the building and refining of simpler, faster production rules. Faster and less complex productions, particularly those that notice internal states, increase the chances of picking up fleeting or intermittent signals related to emotions and epistemic feelings, such as confidence and feelings of knowing (FoK). However, this model does not address the process by which it mitigates emotional reactivity. By extending this research on metacognitive proceduralization, we can investigate a mechanism whereby sufficient metacognitive sensitivity can be developed to deactivate meta-emotions.

Metacognitive skill progresses through stages that are parallel to those within motor skill and cognitive skill, from an early stage of instruction-following to an expert stage that relies on automatic procedural knowledge (production rules).
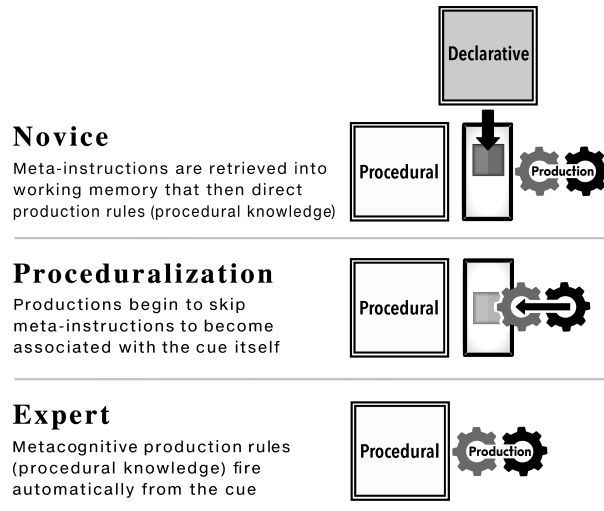
Figure 2: Three stages of metacognitive skill learning through the process of proceduralization (Conway-Smith, West, & Mylopoulos, 2023).

Metacognitive training in detached mindfulness progresses through the following three stages (Figure 2):

The novice stage involves the use of written or verbal meta-instructions to monitor or control some cognitive state (such as attentional training or meditation). In the case of metacognitive training in equanimity, meta-instructions direct the novice's attention toward the momentary changes of affective experience (feeling, sensation, or emotion). These meta-instructions are carried out by productions that retrieve them from declarative memory and execute them. Initial metacognitive performance is slow, effortful, error-prone, and requiring a large degree of working memory.

The intermediate stage of metacognitive training involves the process of proceduralization, where the practice of meta-instructions result in the creation of faster production rules to accomplish the task. Specifically, repeated practice would lead to the compilation of task-specific production rules that bypass declarative knowledge. Because they are faster (due to bypassing declarative memory and possibly being less complex), these productions are more strongly rewarded and more likely to bypass the retrieval of instructions in the future. Metacognitive performance is achieved more quickly, with less effort, and more automatically.

The expert stage involves a robust accumulation of production rules that have been refined and stored in procedural memory. These productions can be deployed automatically to act out monitoring and control processes quickly and effectively. These productions may be faster and less complex, resulting in a lower metacognitive threshold and an improved perception of affective experience. Metacognitive performance in this case demonstrates many characteristics of expertise, i.e., being fast, effective, automatic, and requiring minimal working memory.

Deactivating meta-emotions

Proceduralization, the development of task-specific production rules, assists in providing a computational account of how training to perceive affective variations (equanimity) results in the deactivation of meta-emotions.

Recall that productions rules match and fire off the content of working memory at a default rate of 50ms. That is, productions require at least 50ms to detect a pattern held within working memory. Should a pattern be perceived as sufficiently stable for over 50ms, productions will automatically match and fire off that pattern. Hence, the timing of production rules may be considered a condition of the metacognitive threshold (and psychophysical thresholds more generally) as it provides a partial account of which properties of the stimulus are needed to evoke a response, i.e., strength of signal and perceived stability.

An analogous psychophysical threshold is well known in vision research, where a light that flickers rapidly enough appears to be constant (Landis, 1954). This visual illusion is exploited in film production, where still frames are sped up to 24 frames per second to give images the appearance of consistency. The visual threshold at which still images appear to be constant has been referred to as the "moment of fusion". This visual threshold can be partially raised or lowered due to individual differences such as fatigue and age. For our purposes, the illusion of the flicker-fusion phenomena is comparable to the illusion of affective stability, in that they both rely on a person's inability to perceive change above a certain rate.

Similar to the visual threshold, an individual's metacognitive threshold is variable and can be lowered through attention training to perceive weaker signals from internal cognitive states, such as subtle changes in affect. Proceduralization offers a mechanism for developing and refining production rules that are more sensitive to internal signals, so as to eventually break the illusion of affective consistency.

A key insight into precisely how the refined perception of affective change (equanimity) deactivates emotional reactivity comes from the timing of production rules.

Above the threshold

To the extent that a person's metacognitive threshold is above the 50ms firing rate of production rules, they will perceive any pattern within working memory to be relatively stable. Should a negative emotion appear to be consistent over the 50ms threshold, productions have sufficient time to match and fire a secondary negative emotion in response to the first. Assuming the same conditions, the secondary negative emotion may be perceived and reacted to again, producing a tertiary negative emotion. As long as the metacognitive threshold remains, along with the illusion of affective consistency, production rules may fire automatically and the processes of emotional reactivity may repeat indefinitely.

This explanation sheds light on a potential mechanism that produces the continuous increase in negative emotions as experienced within many psychological disorders. Increasing and persistent cycles of maladaptive emotions are among the most common symptoms of mental illnesses and are associated with Cognitive Attentional Syndrome (CAS; Wells, 2009). A nearly universal phenomenon in cognitive disorders, CAS is a style of negative processing marked by fixed, negatively-biased attention which causes maladaptive emotions to be preserved and heightened, resulting in a continual state of emotional distress.

While there is a lack of computational explanations for the mechanisms underlying this style of maladaptive processing, the timing of production rules can help explain how negatively valenced emotions can be heightened through a process of positive feedback. Production rules also help explain the largely unconscious and involuntary nature of emotional reactions, underscoring the need for metacognitive training to develop productions that counteract them.

Below the threshold

We propose that a key mechanism contributing to the deactivation of meta-emotions is the ability to perceive affective change below the 50ms firing rate of production rules. Reducing the metacognitive threshold below 50ms produces an effect similar to what occurs in the visual flicker-fusion illusion when the speed of the film is reduced below 24 frames per second. The illusion of consistency is broken and one perceives the rapid arising and passing of experience.

This refined perception of affective variations inhibits production rules from matching to the constant fluctuations in working memory (Figure 3). In effect, production rules do not have enough time to identify the rapidly changing pattern of affect. In principle, as long as this metacognitive sensitivity remains, productions are unable to fire secondary emotional reactions.

Lowering one's metacognitive threshold below the 50ms rate requires an expert level of metacognitive skill, as it necessitates the accumulation of production

rules that are sufficiently refined. These expert production rules can better detect the subtle variations in affective experience, and the fleeting signals from other internal cognitive states. Conversely, should one's metacognitive threshold again rise above 50ms, the affective pattern would appear sufficiently stable for emotional reactivity to resume.

This account helps to articulate how the subcomponents of mindfulness training assist in diminishing cycles of negative emotion within psychological disorders such as Cognitive Attentional Syndrome. Individuals who experience CAS are often caught in patterns of negative emotion without a normal exit condition from the informational loop (Wells, 2019). From a computational standpoint, the development of production rules of the type discussed would provide an exit condition from maladaptive emotional loops that would otherwise persist.

This analysis highlights the pivotal role of metacognitive training in emotional regulation, and the key mechanism by which metacognitive practices such as detached mindfulness enhance the ability to perceive emotions without immediately reacting to them.
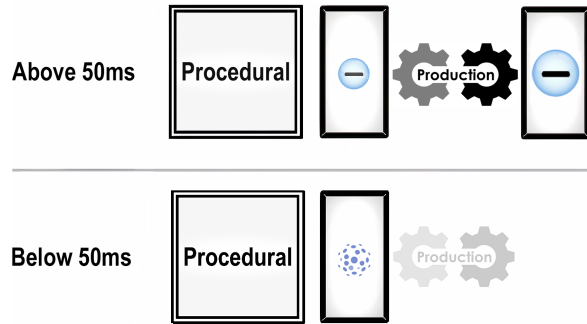
Figure 3. Above the 50ms threshold, an emotion is perceived as sufficiently stable for productions to match and fire secondary emotions. Below the 50ms threshold, the perception of emotional impermanence prevents productions from matching and firing secondary emotions.

Other considerations

Accounting for mindfulness with cognitive modeling is a multifaceted endeavour, and there are many other considerations. For example, there is the issue of buffer decay, or how long patterns of activity can remain within working memory. These issues would apply to representations of both thought and emotion. Another issue is the ability for productions to match to emotional states and to declaratively label them. A particular issue that arises here can be understood in terms of partial matching, or the fidelity of the match. If we take emotion to be a representation of neural activity then we would expect it to have gradations of variability. Since the ability to recognize emotions

would depend on our ability to match to these representational gradients, we would need to assume some form of fuzzy matching. This raises the possibility that some individuals could have more finely tuned productions and conceptual categories for matching emotions, while others may have broader, more fuzzy categories.

Finally, Conway-Smith and West (2023) argued that the capacity of production rules to speed up could increase one's sensitivity to detecting shifts in emotion, and discussed various ways that this speed up could be modeled.

Conclusion

In this paper we have argued that Common Model type architectures can account for important aspects of mindfulness and meditation practices. In particular, we have employed the concept of metacognitive proceduralization to explore the mechanism by which detached mindfulness disengages meta-emotions. A complete model has yet to be constructed, as more theoretical work is required to determine a method of evaluation, considering there is presently no obvious data source with which to compare. One future possibility would be to better articulate the neural correlates of this model and to compare these to the neural imaging results of meditators.

By elucidating the computational processes involved in detached mindfulness and its influence on emotional reactivity, we contribute to a more comprehensive computational understanding that integrates both metacognitive monitoring and control within a unified framework. Meditation on the impermanence of affect is presently an edge case for the Common Model, one that will likely raise questions as to its capacity to simulate it. Our analysis demonstrates that the Common Model framework is able to interpret this practice in a way that accords with reports from practitioners, i.e., the stages of learning, their experiences, and their ability to apply it.

Moreover, by applying the ACT-R cognitive architecture to the study of metacognitive proceduralization, we help bridge the gap between cognitive modeling and psychological practice. The exploration of metacognitive proceduralization within the framework of the Common Model, and specifically ACT-R, offers a novel approach to understanding and intervening in the cycle of negative emotional reactions. Our approach facilitates the exploration of previously underexamined facets of cognitive modeling, aiding in the development of a more complete and integrated cognitive architecture.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological review*.

Anderson, J. R., Betts, S., Bothell, D., Hope, R., & Lebiere, C. (2019). Learning rapid and precise skills. *Psychological review.*

Anderson, J. R., & Fincham, J. M. (2014). Extending problem-solving procedures through reflection. *Cognitive psychology*.

Anderson, J. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Andreu, C. I., Palacios, I., Moënne-Loccoz, C., López, V., Franken, I. H., Cosmelli, D., & Slagter, H. A. (2019). Enhanced response inhibition and reduced midfrontal theta activity in experienced Vipassana meditators. *Scientific reports.*

Baird, B., Mrazek, M. D., Phillips, D. T., & Schooler, J. W. (2014). Domain-specific enhancement of metacognitive ability following meditation training. *Journal of Experimental Psychology.*

Borst, J. P., Nijboer, M., Taatgen, N. A., van Rijn, H., & Anderson, J. R. (2015). Using data-driven model-brain mappings to constrain formal models of cognition. PLoS One, 10(3), e0119673.

Brown, A. L., & DeLoache, J. S. (1978). Skills, plans, and self-regulation. *Children's thinking: What develops.*

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of personality and social psychology*.

Cardaciotto, L., Herbert, J. D., Forman, E. M., Moitra, E., & Farrow, V. (2008). The assessment of present-moment awareness and acceptance: The Philadelphia Mindfulness Scale. *Assessment*, 15.

Charles, L., Chardin, C., & Haggard, P. (2020). Evidence for metacognitive bias in perception of voluntary action. *Cognition*.

Christensen, W., Sutton, J., & McIlwain, D. J. (2016). Cognition in skilled action: Meshed control and the varieties of skill experience. *Mind & Language*.

Conway-Smith, B., & West, R. L. (2023). Metacognitive threshold: a computational account. In *Proceedings of ICCM 2023 - the 21st International Conference on Cognitive Modelling*.

Conway-Smith, B., West, R. L. & Mylopoulos, M. (2023). Metacognitive skill: how it is acquired. In *Proceedings of the Annual Conference of the Cognitive Science Society.*

Dobson, K. S. (2013). The science of CBT: toward a metacognitive model of change?. *Behavior therapy.*

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York: Free Press.

Efklides, A., Schwartz, B. L., & Brown, V. (2017). Motivation and affect in self-regulated learning: does metacognition play a role?. In *Handbook of self-regulation of learning and performance.* Routledge.

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.). *Categories of human learning.* New York: Academic Press.

Fitts, P. M., & Posner, M. I. (1967). *Human performance.* Oxford, England: Brooks/Cole.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist.*

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society.*

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience.*

Fox, K. , Dixon, M. , Nijeboer, S., Girn, M., Floman, J. , Lifshitz, M., & Christoff, K. (2016). Functional neuroanatomy of meditation: A review and meta-analysis of 78 functional neuroimaging investigations. *Neuroscience & Biobehavioral Reviews.*

Grossman, P. (2010). Mindfulness for psychologists: Paying kind attention to the perceptible. *Mindfulness.*

Grossman, C., & Bänninger-Huber, E. (2015). Looking into meta-emotions. *Synthese.*

Hagen, R., Hjemdal, O., Solem, S., Kennair, L. E. O., Nordah, H. M., Fisher, P., & Wells, A. (2017). Metacognitive therapy for depression in adults: A waiting list randomized controlled trial with six months follow-up. *Frontiers in Psychology.*

Hammersmark, A. T., Hjemdal, O., Hannisdal, M., Lending, H. D., Reme, S. E., Hodne, K., ... & Johnson, S. U. (2024). Metacognitive therapy for generalized anxiety disorders in group: A case study. *Journal of Clinical Psychology,* 80(4), 884-899.

Holas, P., & Jankowski, T. (2013). A cognitive perspective on mindfulness. *International Journal of Psychology.*

Kakumanu, R. J., Nair, A. K., Venugopal, R., Sasidharan, A., Ghosh, P. K., John, J. P., ... & Kutty, B. M. (2018). Dissociating meditation proficiency and experience dependent EEG changes during traditional Vipassana meditation practice. *Biological psychology.*

Kim, J. W., & Ritter, F. E. (2015). Learning, forgetting, and relearning for keystroke-and mouse-driven tasks: Relearning is important. *Human–Computer Interaction.*

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *Ai Magazine.*

Landis, C. (1954). Determinants of the critical flicker-fusion threshold. *Physiological Reviews.*

McCormick, C. B. (2003). *Metacognition and learning.* Handbook of psychology.

Newell, A. (1994). *Unified theories of cognition.* Harvard University Press.

Normann, N., & Morina, N. (2018). The efficacy of metacognitive therapy: a systematic review and meta-analysis. *Frontiers in psychology.*

Pauen, M., & Haynes, J. D. (2021). Measuring the mental. *Consciousness and Cognition.*

Pavese, C. (2019). The psychological reality of practical representation. *Philosophical Psychology.*

Pearman, A., Lustig, E., Hughes, M. & Hertzog, C. (2020). Initial evidence for the efficacy of an everyday memory and metacognitive intervention. *Innovation in Aging.*

Posner M. I., Rothbart M. K., Tang Y. (2015). Enhancing attention through training. *Cognitive Enhancement.*

Predatu, R., David, D. O., & Maffei, A. (2020). Beliefs About Emotions, Negative Meta-emotions, and Perceived Emotional Control During an Emotionally Salient Situation in Individuals with Emotional Disorders. *Cognitive Therapy and Research.*

Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness.* OUP Oxford.

Reitter, D. (2010). Metacognition and multiple strategies in a cognitive model of online control. Journal of *Artificial General Intelligence.*

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational psychology review.*

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences.*

Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of cognitive neuroscience.*

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human factors.*

Van Dam, N. T., Van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., ... & Meyer, D. E. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on psychological science.*

Veenman, M. V., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and individual differences.*

Veenman, M., & Elshout, J. J. (1999). Changes in the relation between cognitive and metacognitive skills during the acquisition of expertise. *European journal of psychology of education.*

World Health Organization. (2022). World mental health report: transforming mental health for all. https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf

Wells, A., & Matthews, G. (1994). Self-consciousness and cognitive failures as predictors of coping in stressful episodes. *Cognition & Emotion.*

Wells, A. (2005). Detached mindfulness in cognitive therapy: a metacognitive analysis and ten techniques. J. Ration. Emot. Cogn. Behav. Ther. 23, 337–355.

Wells, A. (2019). Breaking the cybernetic code: Understanding and treating the human metacognitive control system to enhance mental health. *Frontiers in Psychology,* 10, 2621.

West, R. L., & Conway-Smith, B. (2019). Put Feeling into Cognitive Models: A Computational Theory of Feeling. In *Proceedings of ICCM 2019 17th International Conference on Cognitive Modelling.*

Adaptation to Change in Binary Choice: Effects of Interventions and Direction of Change

Maria José Ferreira (mariajor@andrew.cmu.edu)
Department of Social and Decision Sciences
Carnegie Mellon University

Cleotilde Gonzalez (coty@cmu.edu)
Department of Social and Decision Sciences
Carnegie Mellon University

Abstract

Humans show difficulty in adapting to dynamic environments, even in simple scenarios. Researchers have explored different directions to address this issue, including the use of cognitive models to predict human adaptive capabilities. This research investigates the effectiveness of an intervention in helping people improve adaptation to change. We conducted an experiment involving a binary choice task, manipulating the presence of an intervention and the direction of change in outcome payoffs. The change involved two situations: an increasing condition in which one option improves over time and a decreasing condition in which one option deteriorates over time. Our findings reveal that the intervention was effective only in increasing conditions and that human adaptation was better in decreasing conditions. We also construct an Instance-Based Learning (IBL) cognitive model that tracks human behavior and makes one-step-ahead predictions of human decisions. The results of the accuracy of this model's predictions suggest that the IBL model outperforms human participants in adaptation, and it exhibits greater accuracy in predicting humans' choices in Increasing rather than Decreasing conditions. The potential of using IBL model predictions to inform interventions is discussed.

Keywords: Binary task; Dynamic scenarios; Instance Base Learning; Model-Tracing; Adaptation.

Introduction

Binary choice tasks are widely used in foundational research on experiential decision making (Hertwig & Erev, 2009; Erev & Barron, 2005; Gonzalez & Dutt, 2011; Lejarraga, Dutt, & Gonzalez, 2012). Previous research has explored human adaptation to dynamic changes in such tasks, examining scenarios involving limited information about changing probabilities or unawareness of the dynamics of evolving outcomes (Cheyette, Konstantinidis, Harman, & Gonzalez, 2016; Avrahami, Kareev, & Fiedler, 2017; Plonsky & Erev, 2017; McCormick, Cheyette, & Gonzalez, 2022; Konstantinidis, Harman, & Gonzalez, 2022). The findings consistently reveal significant challenges in adapting choices to changing environments. Notably, humans struggle to adjust their choices to situations where options either improve or deteriorate over time. Specifically, research has found that humans exhibit inadequate exploration of possibilities, especially neglecting initially inferior options (McCormick et al., 2022; Konstantinidis et al., 2022).

Research suggests that the presence of full feedback (i.e., providing the forgone outcome in addition to the obtained outcome), can be beneficial when adapting to evolving scenarios (Yechiam & Busemeyer, 2006; Lejarraga & Gonzalez, 2011; Yechiam & Rakow, 2012; Avrahami et al., 2017; McCormick et al., 2022; Konstantinidis et al., 2022). Other research findings suggest that being transparent about the element of change (i.e., showing the value of the outcome that changes over time), can improve human adaptation (McCormick et al., 2022). However, research on ways to improve adaptation is limited.

In this research, we designed an experiment to determine the effect of a subtle but continuous intervention to improve adaptation in changing environments. To replicate the findings of previous research (McCormick et al., 2022; Konstantinidis et al., 2022), we also use two directions of change: an increasing condition in which one option improves over time and a decreasing condition in which one option deteriorates over time. An intervention is expected to provide support in adaptation to change in situations where options improve over time, by encouraging more exploration of the option that may be initially inferior.

We also explore the contribution of cognitive models, particularly those derived from Instance-Based Learning Theory (IBLT Gonzalez, Lerch, and Lebiere (2003)), to understanding human adaptation in these situations (Lejarraga et al., 2012; Konstantinidis et al., 2022). We construct an IBL model to trace human choices and examine the ability of the model to predict one-step-ahead decisions. *Model-tracing* is a technique used to determine the need for feedback in tutoring systems (Anderson, Corbett, Koedinger, & Pelletier, 1995). This technique has been used in IBL models to align the model's memory with that of the human and to be able to make predictions of the human's choices (Cranford et al., 2020; Lebiere et al., 2023). In this paper, we examine the ability of the IBL model with tracing methods to accurately anticipate human sequential decisions one step ahead. We also explore how the intervention and the direction of change manipulations may influence the accuracy of model-tracing. We discuss how the IBL model could inform adaptive interventions.

Experiment: Direction of Change and Interventions

Participants

A sample of 203 participants was recruited from Amazon Mechanical Turk. The ages of the participants ranged from

23 years to 69 years, with 82 females, 118 males, 2 non-binary, and 1 other. Participants were randomly assigned to one of four conditions: Direction of change (increasing and decreasing) and Intervention (present and absent). Decreasing with Intervention ($n = 50$), Decreasing no-Intervention ($n = 54$), Increasing with Intervention ($n = 51$), or Increasing no-Intervention ($n = 48$). Each participant received a base payment of $2 and a bonus payment of 1 cent for every 100 points earned. Bonuses ranged from $1.78 to $3.88.

Design and Procedure

Participants first received a consent form to describe the investigation and the task, followed by a brief demographic questionnaire and the payment rules. They were also required to correctly answer two random attention check questions. Failure to perform the two attention checks would result in their removal from the task without compensation.

The task required participants to make 100 choices from two possible options designated as buttons labeled **A** and **B**. One of the options, stationary (A), gave an outcome of 0 or 500 points with a fixed probability of 50%, while the other option, non-stationary (B), provided outcomes that increased or decreased 10 points per choice in a range of $[10 - 1000]$ or received 0 points with a fixed probability of 50%. The labels of buttons A or B were randomly assigned to the right or left side and remained unchanged thereafter.

In the Increasing condition, the non-stationary outcome starts as 10 points and increases by 10 points per trial to a maximum of 1000 points. In the Decreasing condition, the non-stationary outcome starts with 1000 points and decreases 10 points per trial to a minimum of 10 points. In both these manipulations, the stationary and non-stationary options have the same equivalent expected value over the 100 trials, $EV_{stationary} = EV_{non-stationary} = (0.50 * 0) + (0.50 * 500) = 250$. The non-stationary option's expected value at individual trials depends on the direction of the change. In the Increasing condition, the option with the highest expected value is the stationary one from trials 1 to 49. At trial 50, both options have the same expected value, which we call the switch point. From trials 51 to 100, the non-stationary option has a higher expected value than the stationary option. The decreasing condition has the reverse behavior. We call the best option in each trial the "maximizing" option.

Participants always receive information about the result of the option they select (e.g., "*Your choice was: A and the outcome of your choice was: 500*"). They received additional information on average outcomes of the previous 10 trials in the Intervention conditions (e.g., "*You have selected option A* 10 *times in the previous* 10 *trials and has given you an average of* 500 *points. While you have not selected option B in the past* 10 *trials.*").

At the end of the task, participants answered three questions about their impressions of the option yielding higher points on average for trials 1 to 50 and trials 51 to 100. Participants in the intervention conditions responded to three additional questions about their strategy and their views on the

usefulness of the intervention.

Metrics

The **MaxRate** per participant is the proportion of trials in which a participant chose the maximizing option. MaxRate is calculated per trial as the proportion of participants (out of the total in each condition) who chose the maximizing option. MaxRate is also calculated per block: in block 1, it is the proportion of trials out of 49 ($1 - 49$) in which participants chose the maximizing option, and in block 2 it is the proportion of trials out of 50 (trials $51 - 100$) in which participants chose the maximizing option.

The MaxRate per block helps to determine the level of adaptation (from block 1 to block 2) per participant. We categorize participants into four types of choice behavior based on their **individual adaptation ability**: **Agile** is a maximizer in the first block who continues to maximize in the second, **Clumsy** is a non-maximizer in the first block who shifts to continue to choose the non-maximizing option in the second block, **Fortunate** is a non-maximizer in the first block, who through luck, continues to choose the same option and maximizes in the second block, and **Rigid** is a maximizer in the first block who maintains the same choice, resulting in selecting the non-maximizing option in the second block. Data coding was performed according to the procedure described in (McCormick et al., 2022).

Based on the answers to the post-questionnaire question "*What do you think the relationship was between the two options?*", made for each block separately, participant's **awareness** was coded as: "Aware" if: a) in block 1 Increasing or block 2 Decreasing conditions they answered "*A gave me more points than B, on average*" and b) in block 1 Decreasing or block 2 Increasing conditions they answered "*B gave me more points than A, on average*". The remaining combinations of answers and the participants who replied "*A and B gave me the same number of points, on average*" were classified as "Not-Aware".

Human Experiment Results[1]

MaxRate. Figure 1 shows the MaxRate of the participants in the four conditions on the 100 trials and within the two blocks. The average MaxRate under Decreasing no-Intervention was 64% ($STD = 0.48$) and 61% ($STD = 0.49$) for Decreasing with Intervention conditions. Under Increasing no-Intervention MaxRate was 54% ($STD = 0.50$) and 55% ($STD = 0.50$) for Increasing with Intervention. This result replicates (McCormick et al., 2022; Konstantinidis et al., 2022) in which Increasing conditions result in poorer adaptation compared to Decreasing conditions. Also, visual inspection shows a weak effect of the Intervention in the Increasing conditions.

A generalized logit mixed effects model with the direction of change, Intervention, and block as fixed effects and

---

[1]The statistical analysis, scripts, and dataset for all the results can be found at: https://osf.io/7snra/.
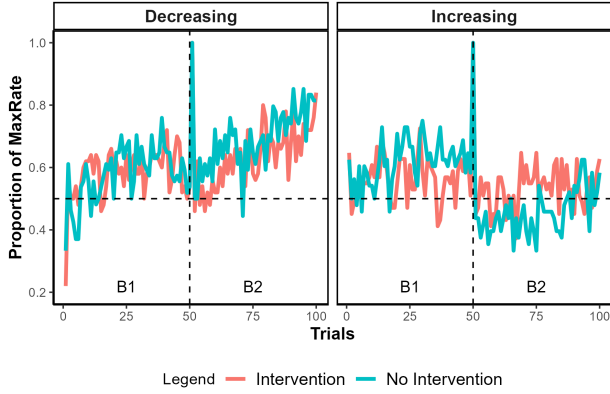
Figure 1: MaxRate performance per trial in each condition, with B1 and B2 denoting the before and after of trial 50.

random intercepts for each participant confirms the expected significant main effect of the direction of change, $\chi^2(1) = 13.81, p < .001$. MaxRates are higher in the Decreasing ($M = 0.62$, $STD = 0.20$) conditions than in the Increasing ($M = 0.54$, $STD = 0.25$) conditions. None of the main effects of Intervention ($\chi^2(1) = 0.10, p = .750$) or block ($\chi^2(1) = 0.07, p = 0.785$) were significant. However, we found a significant two-way interaction between the direction of change and block, $\chi^2(1) = 17.65, p < .001$. In block 1, the MaxRates in the Increasing ($M = 0.59, STD = 0.21$) and Decreasing conditions ($M = 0.58, STD = 0.19$) are not statistically different $t(398) = -0.34, p > .050^2$. In contrast, in block 2, the MaxRates are higher in Decreasing conditions ($M = 0.67, STD = 0.21$) than in Increasing conditions ($M = 0.49, STD = 0.27$), $t(398) = 5.60, p < .001$.

The two-way interactions between the Intervention and the direction of change ($\chi^2(1) = 0.70, p = .402$) and between the Intervention and block ($\chi^2(1) = 0.64, p = .425$) were not significant. However, there was a significant three-way interaction between Intervention, direction of change, and block $\chi^2(1) = 4.31, p = .038$. The MaxRates in block 1 did not show significant differences (all $p > .050$) regardless of the presence or absence of an Intervention; but there was a significant difference in MaxRates between Decreasing ($M = 0.69, SD = 0.20$) and Increasing ($M = 0.46, SD = 0.27$) conditions, in block 2 when the Intervention was absent, $t(398) = 5.43, p < .001$ and no significant differences when the Intervention was present, $t(398) = 2.50, p = .155$: Decreasing ($M = 0.64, SD = 0.21$) and Increasing ($M = 0.53, SD = 0.27$). This result suggests that the presence of the Intervention caused the Increasing condition to exhibit higher maximization, comparable to the Decreasing with the Intervention condition, implying a relative benefit of the Intervention in the Increasing condition.

Individual Adaptation. Table 1 reports the proportions of Agile, Clumsy, Fortunate, and Rigid participants. It also

<hr>

[2]The Bonferroni correction was applied to adjust all p-values when examining interaction effects to account for multiple tests.

presents summaries of Adaptive and Non-adaptive proportions and the differences between Decreasing and Increasing conditions and between the presence and absence of an Intervention.

These analyses indicate a larger proportion of adaptive participants in Decreasing compared to Increasing conditions. Approximately half of the participants overall were adaptive, and most of the effects of adaptation are due to the greater proportion of adaptive participants in the Decreasing than the Increasing conditions. It also shows a general ineffectiveness of the Intervention. Figure 2 provides a more nuanced view. It shows a slight positive effect of the Intervention in the Increasing conditions, where the proportion of Agile and Fortunate type of participants is larger in the Increasing condition with Intervention compared to the Increasing condition without intervention. In other words, the Intervention helped some participants in the Increasing condition maximize in the second block. Some were fortunate as they continued to choose the same option they were choosing in the first block, and others intentionally shifted to the maximizing option in the second block (Agile).
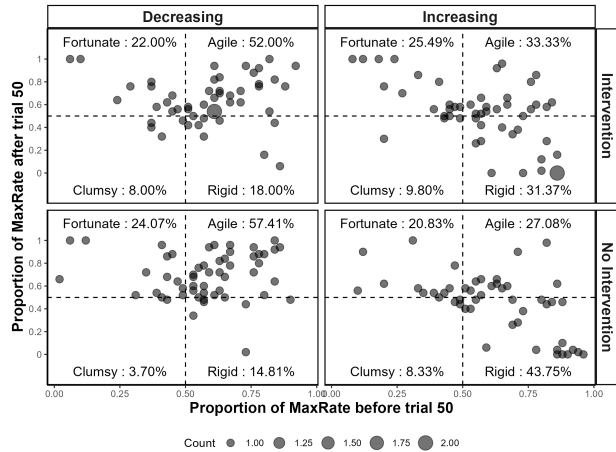
Figure 2: Participant's behavior split into four quadrants denoted by the dash lines at 0.5 using their individual MaxRate performances before and after trial 50.

Awareness. The proportion of Aware participants per condition suggests that more than 50% of the participants were explicitly aware of the change in the choice environment: 72% in the Decreasing Intervention, 67% in the Decreasing no-Intervention, 56% in the Increasing no-Intervention, and 55% in the Increasing Intervention conditions. These patterns agree with the MaxRates findings regarding the effect of the direction of change. Just like in the MaxRate results, individuals also report being more aware of the changes in the Decreasing conditions than in the Increasing conditions, and the Intervention did not affect such awareness.

The proportion of MaxRates of block 2 for Aware and Not-Aware participants was analyzed using 4 Mann-Whitney U tests. The results show significant differences among Aware and Not-Aware participants. The Increasing condi-

Table 1: Percentage of maximization pattern based on individual differences for participants in each condition.

Maximization Pattern	Averaged over Direction of Change			Averaged over Intervention			All
	Inc.	Dec.	Difference (Dec. - Inc.)	Int.	No Int.	Difference (Int. - No Int.)	
Agile	30.30%	54.81%	24.51%	42.57%	43.14%	-0.57%	42.86%
Clumsy	9.09%	5.77%	-3.32%	8.91%	5.88%	3.03%	7.39%
Total Adaptive	**39.39%**	**60.58%**	**21.19%**	**51.49%**	**49.02%**	**2.47%**	**50.25%**
Fortunate	23.23%	23.08%	-0.15%	23.76%	22.55%	1.21%	23.15%
Rigid	37.37%	16.35%	-21.02%	24.75%	28.43%	-3.68%	26.60%
Total Non-Adaptive	**60.61%**	**39.42%**	**-21.17%**	**48.51%**	**50.98%**	**-2.47%**	**49.75%**
Difference (Adaptive - Non-Adaptive)	-21.22%	21.16%		2.97%	-1.96%		0.50%

tions had higher MaxRates for Aware ($M = 0.64, STD = 0.18$) than Not-Aware ($M = 0.31, STD = 0.25$) participants, $U = 2080, p < 0.01$; similarly to the Decreasing conditions (MaxRates for Aware: $M = 0.73, STD = 0.17$ and Not-Aware participants: $M = 0.54, STD = 0.23$), $U = 1741, p < 0.01$. The Intervention conditions had higher MaxRates for Aware participants ($M = 0.70, STD = 0.18$) than Not-Aware participants ($M_= 0.39, STD = 0.25$), $U = 1995.5, p < 0.01$; similarly to the no-Intervention conditions (MaxRates for Aware: $M = 0.68, STD = 0.18$ and Not-Aware participants: $M = 0.42, STD = 0.29$), $U = 1921, p < 0.01$.

Instance-Based Learning Model and Model-Tracing

An IBL binary choice model similar to that of previous research (Gonzalez & Dutt, 2011; Lejarraga et al., 2012; Lejarraga, Lejarraga, & Gonzalez, 2014; Konstantinidis et al., 2022) was implemented using PyIBL (Morrison & Gonzalez, 2023). In IBLT, a choice occurs by activating memories of past experiences (e.g., observed outcomes) associated with each option/decision. Memory activation is modulated by at least two processes (i.e., free parameters in the model): memory decay and noise associated with the retrieval of these memories. The activation of outcome i in each option j on trial t is illustrated in the following equation (the complete activation equation in ACT-R is in (Anderson & Lebiere, 1998)):

$$A_{j,i,t} = \sigma \ln \left(\frac{1 - \gamma_{j,i,t}}{\gamma_{j,i,t}} \right) + \ln \sum_{t_p \in \{1,...,t-1\}} (t - t_p)^{-d} \quad (1)$$

where d is a decay parameter, σ is a noise parameter, $\gamma_{j,i,t}$ is a random sample of a uniform distribution (between 0 and 1), and t_p denotes all previous trials in which the outcome i was observed. Past research has found that successful adaptation to choice environments is associated with higher levels of decay d, (Konstantinidis et al., 2022), suggesting that poorer memory leads to better adaptation.

The activation of each instance in memory determines the probability of it being retrieved. This probability is calcu-

lated for each instance i on the activations of all the outcomes observed in each option j:

$$P_{j,i,t} = \frac{e^{A_{i,t}/\tau}}{\sum_j e^{A_{j,t}/\tau}} \quad (2)$$

where τ is a temperature parameter that controls the uniformity of the probability distribution defined by this soft-max equation with a default value of $\tau = \sigma\sqrt{2}$. Finally, the model chooses the option with the highest *blended* value V:

$$V_{j,t} = \sum_{i=1}^{n} P_{i,t} x_i \quad (3)$$

where x is the value of the observed outcome i from option j, P is the probability of retrieval of this outcome as defined in Equation 2, and n is the number of unique outcomes in option j.

As suggested in previous research (Lejarraga et al., 2012), we created initial expectations using a prepopulation of memory instances, using 125 in the Increasing and 625 as outcomes in the Decreasing conditions, to allow a fair start for the stationary options. We also allowed the model to experience the two options, with the maximizing option always being the first, and both reinforced with a low-value outcome of 0. After this point, the model starts making predictions according to the blended value (Equation 3).

Model Fitting and Model-Tracing Procedure

We estimated the best decay parameter by running model-tracing for each participant using each of the 291 decay values within a range of $[0.1 - 3]$ with increments of 0.01. The model tracing was performed with noise $\sigma = 0$ and $\tau = 1$.

The value of decay that corresponded to the lowest Root Mean Square Error (RMSE) between user decisions and model predictions was selected as the "best fitting" decay parameter for that participant. In the case of multiple RMSE with the lowest value, the highest decay was chosen.

To run model-tracing, the IBL agent will have in memory the exact past decisions of the participant. We remove the noise parameter ($\sigma = 0$) and set $\tau = 1$. In this way, the model

will not produce noisy activations (Equation 1). We used the best decay value that fits each individual's data.

For each participant, we created an IBL agent using Py-IBL (Morrison & Gonzalez, 2023) as previously described. When the model starts making decisions, for each new prediction it makes in each trial, we strengthen the decision experienced by the participant. In practice, we replace the instance corresponding to the prediction made by the model with an instance that contains the decision and the outcome experienced by the participant. This process allows us to predict the next-trial decision and adjust the model's memory with the human's exact experience.

Results

Delta MaxRate. To explore the maximization decisions of the model and human, we analyze the difference in MaxRates between the model and the human, called *Delta MaxRate*.

Figure 3 presents the aggregated Delta MaxRate per trial (without trial 1 since the model's choice is random). If a model is able to predict the human maximization choices, the Delta MaxRate would be around the zero line in both blocks. However, although close to the zero line, we observe some deviations between the model choice predictions and human choices.
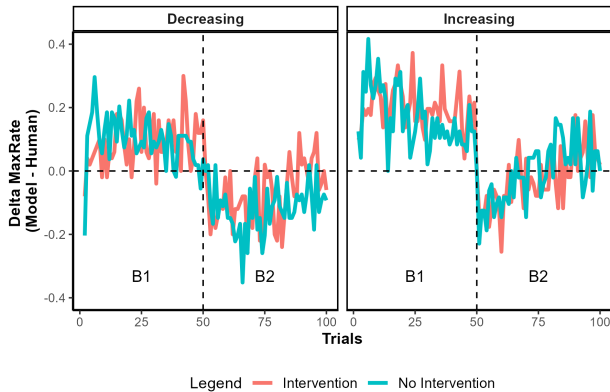
Figure 3: Proportion of Delta MaxRate per condition and over trials (from trials 2 to 100.)

We observe that the MaxRate predicted by the models is larger than the actual human MaxRate in block 1. In the Increasing Intervention condition, Delta MaxRate was $M = 0.20, STD = 0.50$, and $M = 0.17, STD = 0.50$ in the Increasing no-Intervention condition. In the Decreasing Intervention condition, Delta MaxRate was $M = 0.09, STD = 0.55$, and $M = 0.10, STD = 0.55$ in the Decreasing no-Intervention condition. In other words, the model predicted the human maximization decisions more accurately in the Decreasing than Increasing conditions in Block 1.

In contrast, the MaxRate predicted by the models is lower than the actual human MaxRate in block 2, but closer to zero. In the Decreasing Intervention condition, Delta MaxRate was $M = -0.07, STD = 0.58$, and $M = -0.12, STD = 0.58$ for Decreasing no-Intervention. In the Increasing Intervention

condition, Delta MaxRate was $M = -0.02, STD = 0.49$, and $M = -0.01, STD = 0.48$ in the Increasing no-Intervention condition. That is, the model was better than humans in Block 1 but adapted more poorly than humans in block 2, especially in the Decreasing conditions.

Delta MaxRate for Individual Adaptation. Figure 4 presents the Delta MaxRate in blocks 1 and 2 for each participant in our study. Participants who were not accurately predicted by the model (i.e., the Delta MaxRate was greater than or equal to 0.1) are reported in each of the quadrants as an orange dot (close to zero *No*), while participants who were correctly predicted (i.e., the Delta MaxRate was in the range $[0; 0.09]$) are reported as a green dot (close to zero *Yes*).
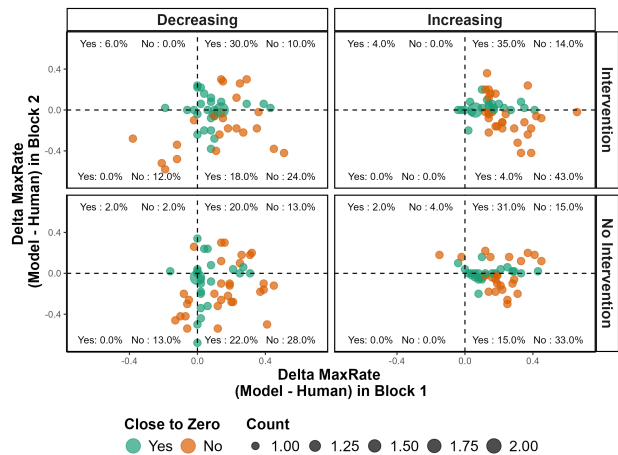
Figure 4: Proportion of Delta MaxRate (Model - Human) in block 1 (X-axis) and block 2 (Y-axis) for each participant.

As is evident in the top-right quadrants, the model demonstrates greater accuracy in predicting human MaxRate under Increasing conditions compared to Decreasing conditions, especially under no Intervention. For example, under Increasing conditions, the model improved the correct anticipation of maximization choices from 31% in the no-Intervention condition to 35% in the Intervention condition, while under Decreasing conditions, the model improved the correct anticipation of maximization choices from 20% in the no-Intervention condition to 30% in the Intervention condition.

Conversely, in the bottom-right quadrants, the model exhibits reduced accuracy in predicting maximization behavior for block 2. For example, under Increasing conditions, the model correctly anticipates a smaller proportion of participants (4% and 15% for the presence and absence of Intervention, respectively), while under Decreasing conditions, this accuracy is higher (18% and 22% for the presence and absence of Intervention, respectively).

Model Synchronization per Trial. To explore more precisely how the model predicts each of the human choices (regardless of whether they were maximizing choices or not), we examine the synchronization between the model prediction and each of the human choices. To calculate this one-step-

ahead prediction of the model, for each participant in each trial[3] t, we determined whether the model prediction was the same as the actual human action. If it was the same, the synchronization value was 1; otherwise it was 0. We aggregated Synchronization Rates (SyncRate) at various levels: averaging the SyncRate per participant (i.e., the average synchronization out of 99 trials), per trial, per block, or per condition. Figure 5 shows the average SyncRate of all participants in each trial by condition.
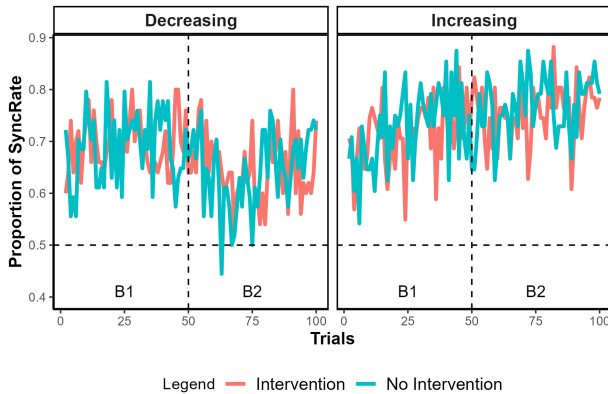
Figure 5: Proportion of synchronization between model and human per condition and over trials (from trials 2 to 100.)

As observed in Figure 5, the model synchronized with human actions above chance. It was able to consistently increase synchronization with human actions over the course of the trials in the Increasing conditions, but synchronization suffered slightly in the second block of the Decreasing condition. Generally, the SyncRate was higher in the Increasing conditions (75%, $STD = 0.14$ for Increasing no-Intervention condition, 73%, $STD = 0.12$ for Increasing with Intervention) than in the Decreasing conditions (66% for both Decreasing conditions, $STD_I = 0.13$ and $STD_{NI} = 0.10$). An analysis of SyncRates using a Kruskal-Wallis test revealed significant differences between the conditions, $H(3) = 18.39, p < 0.01$. Post hoc tests using Bonferroni corrections revealed significant differences between Increasing and Decreasing in the absence of an Intervention ($p = 0.008$) and Increasing and Decreasing in the presence of an Intervention ($p = 0.032$).

Discussion

How humans adapt their choices in dynamic environments to maintain the best outcomes is an unresolved dilemma. This paper investigates two factors that influence this choice adaptation: the direction of change and the presence of continuous interventions. We also investigated whether computational IBL models can accurately capture human adaptive behaviors. Consistent with previous studies (McCormick et al., 2022; Konstantinidis et al., 2022), participants in Decreasing conditions showed better adaptation than in Increasing conditions. This study augments these results by finding that a

---
[3]Except for **trial=1**, since the model selects an option randomly and it cannot accurately predict the participant's decision.

continuous intervention can help participants become more agile and less rigid in the Increasing conditions, although it did not influence the Decreasing conditions.

The lack of effect of the Intervention on improving the adaptation of choices in the Decreasing conditions needs to be further investigated. There are two reasonable explanations for this result. First, it could be argued that the Intervention did not improve people's awareness of the change. This interpretation is supported by our awareness results, which indicate that participants who were aware of the change performed better, regardless of whether they received an intervention. This finding is also consistent with previous research of (Lejarraga & Gonzalez, 2011), in which decision-makers have been found to neglect the descriptive information displayed continuously after having had experience with the task. Second, it is possible that participants felt overwhelmed by the Intervention and the partial feedback received in each trial. This, in turn, could have impaired their ability to process informed judgments continuously. This explanation is consistent with the research of (Gonzalez, 2005), in which the authors found that people are often overwhelmed by processing more information as feedback while making decisions in a dynamic task.

Although modest, the positive impact of the Intervention on the Increasing conditions suggests potential benefits for participants struggling with adapting to rising trends. However, further research is warranted. Specifically, future studies should explore the frequency and timing of the Intervention and trigger conditions to enhance each participant's adaptive behavior. Future research should also explore the form or content of the Intervention message.

As an immediate step, we plan to use our cognitive model and its model-tracing capabilities to determine the frequency and timing of the Intervention. The synchronization of the model's choices with human choices is an encouraging development. Notably, the model consistently improves its synchronization with human participants and is able to improve its maximization predictions under Increasing conditions.

Using model-tracing, we plan to identify participants who are not maximizing as predicted by the model and trigger the Intervention accordingly. Future studies will focus on developing methods to use the model synchronization to decide on the time for intervention. Specifically, we can evaluate the SyncRate values for each trial or group of trials to determine when the model outperforms the user, allowing us to initiate interventions when the model performs better than the human. We will also investigate the missed synchronization cases of the model. For example, in our model, the tau parameter adds noise to the choice predictions, and individually adjusting this parameter can provide a higher model synchronization rate.

Acknowledgments

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167-207. doi: 10.1207/s15327809jls0402_2

Anderson, J. R., & Lebiere, C. J. (1998). *The atomic components of thought (1st ed.).* Psychology Press. doi: https://doi.org/10.4324/9781315805696

Avrahami, J., Kareev, Y., & Fiedler, K. (2017, January). The dynamics of choice in a changing world: Effects of full and partial feedback. *Memory & Cognition*, *45*(1), 1–11. doi: 10.3758/s13421-016-0637-4

Cheyette, S. J., Konstantinidis, E., Harman, J. L., & Gonzalez, C. (2016). Choice adaptation to increasing and decreasing event probabilities. *Cognitive Science*.

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., & Lebiere, C. (2020). Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science*, *12*(3), 992-1011. doi: https://doi.org/10.1111/tops.12513

Erev, I., & Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psychological Review*, *112*(4), 912–931. (Place: US Publisher: American Psychological Association) doi: 10.1037/0033-295X.112.4.912

Gonzalez, C. (2005). Decision support for real-time, dynamic decision-making tasks. *Organizational Behavior and Human Decision Processes*, *96*(2), 142-154. doi: https://doi.org/10.1016/j.obhdp.2004.11.002

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review*, *118*(4), 523.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517-523. doi: https://doi.org/10.1016/j.tics.2009.09.004

Konstantinidis, E., Harman, J. L., & Gonzalez, C. (2022). Patterns of choice adaptation in dynamic risky environments. *Memory & Cognition*, *50*(4), 864–881. doi: 10.3758/s13421-021-01244-4

Lebiere, C., Cranford, E. A., Aggarwal, P., Cooney, S., Tambe, M., & Gonzalez, C. (2023). Cognitive modeling for personalized, adaptive signaling for cyber deception. In T. Bao, M. Tambe, & C. Wang (Eds.), *Cyber deception: Techniques, strategies, and human aspects* (pp. 59–82). Cham: Springer International Publishing. doi: 10.1007/978-3-031-16613-6_4

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes*, *116*(2), 286-295. doi: https://doi.org/10.1016/j.obhdp.2011.05.001

Lejarraga, T., Lejarraga, J., & Gonzalez, C. (2014). Decisions from experience: How groups and individuals adapt to change. *Memory & Cognition*, *42*(8), 1384–1397. doi: 10.3758/s13421-014-0445-7

McCormick, E. N., Cheyette, S. J., & Gonzalez, C. (2022). Choice adaptation to changing environments: trends, feedback, and observability of change. *Memory & Cognition*, *50*(7), 1486–1512. Retrieved from https://doi.org/10.3758/s13421-022-01313-2 doi: 10.3758/s13421-022-01313-2

Morrison, D., & Gonzalez, C. (2023). *PyIBL: A python implementation of IBLT.* http://pyibl.ddmlab.com. (Version 5.0.3)

Plonsky, O., & Erev, I. (2017). Learning in settings with partial feedback and the wavy recency effect of rare events. *Cognitive Psychology*, *93*, 18-43. doi: https://doi.org/10.1016/j.cogpsych.2017.01.002

Yechiam, E., & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, *19*(1), 1-16. doi: https://doi.org/10.1002/bdm.509

Yechiam, E., & Rakow, T. (2012). The effect of foregone outcomes on choices from experience. *Experimental Psychology*, *59*(2), 55-67. (PMID: 21914593) doi: 10.1027/1618-3169/a000126

Integrating Social Sampling Theory into ACT-R: A Memory-Based Account of Social Judgment and Influence

Christopher R. Fisher (christopher.fisher.27.ctr@us.af.mil)
Parallax Advanced Research
Beavercreek, OH 45324 USA

Taylor Curley (taylor.curley@us.af.mil)
Air Force Research Laboratory
Wright Patterson AFB, OH USA

Abstract

Cognitive architectures (CAs) have been instrumental in integrating a wide range of findings in cognitive science into unified theories of cognition. However, much less effort has been devoted to applying CAs to social phenomena, despite the high interdependence between cognitive and social processes in real-world scenarios (e.g., Ecker et al., 2022). We integrated social sampling theory (SST) and ACT-R to begin filling this gap. ACT-R is a modular, hybrid symbolic/sub-symbolic CA with a detailed memory system. SST describes how beliefs and behavior emerge from an interplay between individual and social motivations. The component theories have complementary strengths and weaknesses: SST provides an account of social influence and comparison, but lacks a memory system to support those processes, whereas the converse is true for ACT-R. In two simulations, we demonstrate that SST-ACT-R produces social influence dynamics not present in either component theory. Specifically, SST-ACT-R shows how private and publicly expressed beliefs may evolve through social interactions based on social influence and underlying memory mechanisms.

Keywords: ACT-R; social sampling theory; memory; social cognition; agent-based modeling

Introduction

Cognitive architectures (CAs) are computational frameworks for simulating and evaluating unified theories of cognition (Newell, 1990). What separates a CA from other theories is the focus on identifying invariant properties in the structure and function cognition, which applies to phenomena across a wide variety of domains. Considerable progress has been made in developing CAs which account for phenomena in domains as diverse as memory (Anderson et al., 1998), visual search (Nyamsuren & Taatgen, 2013), problem solving (Anderson et al., 2004), and decision making (Gonzalez et al., 2003), among others.

Much less research has been devoted to using CAs to model and explain phenomena in social psychology. Applying CAs to the domain of social psychology is important for two reasons: (1) as a general theories of cognition, CAs should generalize to social behavior, and (2) decades of research show that cognition and behavior do not occur in a social vacuum, but instead are often moderated by social context. Some examples include *polarization*, which occurs when beliefs become more extreme after members of a group communicate with each other (Lord et al., 1979), and the *false consensus effect*, which occurs when a person overestimates how widely his or her own beliefs are held by others (Ross et al., 1977). In more recent years, the real-world impact of social influence has been observed in the increased spread of misinformation regarding political news (Allcott et al., 2019) and public health information (Suarez-Lledo & Alvarez-Galvez, 2021) through social media. A comprehensive understanding of the spread and influence of misinformation likely requires both models of cognition and social influence (Ecker et al., 2022).

Some notable, albeit limited, efforts have been made to apply CAs to the domain of social psychology. For example, Stevens et al. (2016) developed meta-cognitive agents based on ACT-R which interacted in the Prisoners' Dilemma. Each agent had a theory of mind, allowing it to reason about knowledge and strategies used by the other agent. Additionally, the CLARION CA has been used to model organizational decision making behavior and the development of collaborations in the academic publication process (Sun, 2007). In both cases, the success of the models were attributed to increased cognitive realism, such as including learning and decision processes similar to those of real humans. In one last example, researchers instantiated Festinger's social comparison theory into the SOAR CA to model the emergence of imitation in crowd behavior (Fridman & Kaminka, 2011). Nonetheless, many gaps exist in the literature bridging CAs and social psychology. For example, it is unclear how a CA such as Adaptive Control of Thought-Rational (ACT-R; Anderson et al., 2004) can make relative social comparisons, and moderate publicly expressed beliefs based on social context.

Our goal is to expand CAs further into the domain of social psychology by integrating Social Sampling Theory (SST; Brown et al., 2022) and the CA, ACT-R (Anderson et al., 2004). An integrative approach has the potential to create a more comprehensive CA resulting from the complementary strengths and weaknesses of ACT-R and SST. On one hand, SST is a theory of social comparison and influence which explains a wide variety of social phenomena, such as polarization, and the false consensus effect (Brown et al., 2022). However, SST does not specify the memory processes which give rise to its social comparison process. Instead, SST hard-codes static belief distributions into the model without explaining how they are represented and potentially change across time with experience. On the other hand, the memory system of ACT-R is well-specified and has endured many decades of empirical testing (Anderson et al., 2004, 1998). However, ACT-R does not specify the details of social influence and comparison processes.

Overview

The remainder of the paper is structured as follows. In the next two sections, we introduce SST and ACT-R and their

core tenets. With that foundation laid, we then provide a justification for integrating SST and ACT-R and describe the details of the integration. Next, we describe two simulations which demonstrate new behaviors arising from the integration of SST and ACT-R: (1) stochasticity and learning/forgetting dynamics in belief distributions, and (2) the evolution of belief distributions through social interactions. Finally, we conclude with a discussion of limitations.

Social Sampling Theory

Social Sampling Theory (SST) is a social-cognitive theory of social judgment and social influence which explains a wide variety of social phenomena, including, but not limited to, false consensus, polarization, the backfire effect, and social contagion (Brown et al., 2022). SST is predicated on the following assumptions: (1) private beliefs and social norms are represented as distributions rather than a single values, (2) social comparisons are based on the ranking of a belief within a distribution, (3) social comparisons are based on the retrieval of small samples from memory of the immediate social environment, and (4) when expressing a belief publicly, individuals strive to find a compromise between two potentially competing forces—*authenticity preference*, and *social extremeness aversion* (Brown et al., 2022; Kuran, 1997). Authenticity preference refers to the desire to publicly express beliefs that are consistent with one's privately held beliefs. Social extremeness aversion refers to the desire to not deviate from a social norm. A social norm is typically based on the immediate social situation.

Figure 1 illustrates the process of selecting a belief to publicly express. Following Brown et al. (2022), we represent beliefs as beta distributions on a continuum ranging from 0 (liberal) to 1 (conservative). Note that other distributions and beliefs could be expressed within SST. The private belief distribution in red is diffuse and shifted to the left, reflecting the fact that the private beliefs tend to be liberal, but are weakly held. By contrast, the social norm distribution is shifted to the right and more concentrated, indicating that expressed beliefs of the social group are consistently conservative. A balance between authenticity preference and social extremeness avoidance is accomplished by evaluating the rank of a candidate belief relative to each distribution using the median (i.e., the least extreme rank) as a reference point.

In SST, social comparisons are based on the rank of a belief within a distribution, which implies that a belief distribution with a smaller variance will contribute more to the cost function than one with a larger variance. In Figure 1, the cost of deviating from the median is much less for the private belief distribution compared to the social norm distribution because the private beliefs are more diffuse. To understand why, consider a point that is .1 units above the median of the wide, private belief distribution and .1 units below the median of the narrow, social norm distribution. The percentile of this point for the private belief distribution is moderate at 63%, whereas the percentile for the social norm distribution is much more
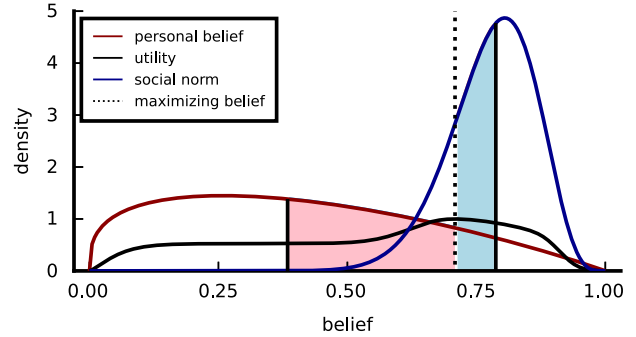
Figure 1: An illustration of selecting the utility-maximizing belief to publicly express using $w = .50$, $\gamma = 20.0$, and $\delta = 1.0$. Solid vertical black lines are medians. Utility was rescaled to the range $[0,1]$.

extreme at 14%.

The publicly expressed belief is determined by maximizing a utility function which takes into consideration the ranking of a belief relative to each distribution. The solid black curve in Figure 1 shows the utility as a function of belief, with the maximum located at the dotted vertical, black line. The deviation from median of the private belief and social norm distributions are represented by the pink and light blue areas, respectively. As expected, the utility maximizing belief is pulled closer to the median of the social norm distribution due to its lower variance.

More formally, let $F(x; \alpha_z, \beta_z)$ be the cumulative distribution function (CDF) of the beta distribution with parameters α and β and index $z \in \{p, s\}$, which represents the private belief and social norm respectively. Let $H_z(x) = |.50 - F(x; \alpha_z, \beta_z)|$ be the area between the median and belief x. Because $H_z(x)$ represents a cost, the terms in the equation below are multiplied by -1 to represent utilities. The utility function is defined as:

$$U(x) = -we^{\gamma \cdot H_s(x)} - (1-w)e^{\gamma \cdot H_p(x)}, \quad (1)$$

where $w \in [0,1]$ is the weight placed on the social norm, and $\gamma \geq 0$ represents sensitivity to increases in the deviation from the median. The utility maximizing belief is defined as:

$$b_{\max} = \arg\max_{x \in [0,1]} U(x), \quad (2)$$

reflecting the weighting of an agent's private belief and the observed social norm.

ACT-R

Adaptive Control of Thought-Rational (ACT-R) is a hybrid symbolic, non-symbolic CA which has been applied to many domains including memory (Anderson et al., 2004) and decision making (Gonzalez et al., 2003). Unlike other modeling approaches, the goal of CAs is to provide a unified theory of cognition, spanning areas as diverse as perception, action,

memory, and decision making (Newell, 1990). ACT-R's architecture is structured as a set of specialized information processing units called modules. The set of modules includes capabilities for visual and auditory perception, declarative and procedural memory, and goal maintenance. The procedural module coordinates activity of the other modules using condition-action pairs called production rules (Anderson et al., 2004). During each production cycle, the procedural module selects a production rule based on its match to the state of the architecture and then executes the actions.

For our current purposes, it is sufficient to focus on the declarative memory module. ACT-R's declarative memory module is responsible for encoding, storing, and retrieving factual information which can be verbally expressed. Declarative memory consists of set of chunks $\mathbf{C} = \{\mathbf{c}_m\}_{m \in I_C}$ where I_C is an index set. A chunk considered the basic unit of declarative knowledge, and consists of a set of slot-value pairs. A chunk is defined as:

$$\mathbf{c}_m = \left\{(s_{m,i}, v_{m,i})\right\}_{i \in I_{c_m}}, \tag{3}$$

where $s_{m,i}$ and $v_{m,i}$ are the slot and value of pair i. A concrete example of a chunk is $\mathbf{c} = \{(\text{object}, \text{house}), (\text{color}, \text{blue})\}$, which corresponds to a blue house. The value associated with a slot can be queried with the function $c_m(s) = v$, which is useful for specifying retrieval requests, $\mathbf{r}$. In the present model, a chunk consists of the following slots $Q = \{\text{source,issue,id,belief}\}$. The slot *source* indicates whether a belief is private or a social norm. The slot *issue* corresponds to an issue or topic index (e.g., renewable energy). The slot *id* corresponds to the unique index of the agent associated with the belief. The slot *belief* encodes a belief with possible values ranging from 0 (maximally liberal) to 1 (maximally conservative).

Each chunk is associated with an activation value, representing its ability to be retrieved (Anderson et al., 2004). Increasing activation increases the probability and speed with which a chunk can be retrieved activation. In the present model, activation is defined as:

$$a_m = B_m + \rho_m + \varepsilon_m, \tag{4}$$

where B_m is base-level learning, ρ_m the partial matching term, and $\varepsilon_m \sim \text{logistic}(0, \eta)$ is activation noise with scalar parameter η. Base-level learning governs the dynamics of learning from experience, and forgetting across time, and is defined as:

$$B_m = \log\left(\sum_{j=1}^{n_m} t_{m,j}^{-d}\right), \tag{5}$$

where n_m is the number of times chunk m has been used or retrieved, $t_{m,j}$ is elapsed time in seconds since the jth retrieval, and $d \in [0, 1]$ is a decay parameter. Partial matching controls discriminability between the retrieval request $\mathbf{r}$ and chunk $\mathbf{c}_m$, and is given by:

$$\rho_m = -\delta \sum_{q \in Q} I(c_m(q), r(q)), \tag{6}$$

where Q is the set of slot values in the retrieval request, $\delta \geq 0$, and I is an indicator function, which returns 1 in the case of a mismatch, and 0 otherwise.

ACT-R uses the blending mechanism from instance based learning (IBL; Gonzalez et al., 2003) to estimate the expected value of a slot-value over chunks, $\{v_{m,k}\}_{m \in I_C}$. The estimate is called a blended value and is computed as a weighted average in which the probability of retrieving a given chunk serves as a weight. The blended value for slot-value k is defined as:

$$\widehat{\mu_{b_k}} = \sum_{m | \mathbf{c}_m \in \mathbf{C}} p_m v_{m,k}, \tag{7}$$

where $v_{m,k}$ is the value of slot k in chunk m, and p_m is the probability of retrieving chunk m. The retrieval probability is given by the softmax function:

$$p_m = \frac{e^{a_m/\tau}}{\sum_{j | \mathbf{c}_j \in \mathbf{C}} e^{a_j/\tau}}, \tag{8}$$

where τ controls how sensitive the probability weights are to activation.

Integration

A close comparison of ACT-R and SST reveals complementary strengths and weaknesses. An advantage of ACT-R is its well-specified and validated declarative memory system (Anderson et al., 1998), which describes memory representation, the process of memory retrieval, and the dynamics of learning and forgetting. One limitation of ACT-R, however, is that it lacks cognitive mechanisms that are sensitive to social influence and context. The opposite is true for SST: it provides an account of the social comparison process and a utility function underlying decision making, but lacks an account of the supporting memory mechanisms. Instead, the belief distributions in SST are hard-coded by the modeler rather than emerging from more basic principles. By integrating SST into ACT-R, it is possible to capitalize on their strengths to describe how learning through social interactions creates dynamics in belief distributions.

The basic integration of SST into ACT-R is straightforward: rather than assuming specific belief distributions, the distributions are based information encoded in ACT-R's declarative memory system, subject to the dynamics described in Equation 4. One important component of the integration is deciding how the rank of a belief within a distribution is determined. We discuss two approaches. One approach evaluates the percentiles from an empirical CDF, $\widehat{F(x)}$, which is evaluated against weighted slot-values, $\{p_m v_{m,k}\}_{m \in I_C}$. One advantage of this approach is that it does not rely on parametric assumptions regarding the distribution of slot-values. However, percentile estimates will be coarse when the number of contributing chunks is small. In some cases, this issue can be partially mitigated by incorporating background knowledge in the form of chunks—a kind of

pseudo prior—to improve estimates of the percentiles. Precedence for this approach can be found in numerous applications of IBL (e.g., Gonzalez & Dutt, 2011), where declarative memory is initialized with a chunk to provide the model with a means of responding on the first trial. An alternative approach—which we adopt here for simplicity—is to fit a parametric distribution to the slot-values to improve the resolution of the percentiles. For simplicity, we reparameterize the beta distribution in terms of a mean and standard deviation. The variance is the expected value of the squared difference between a random variable and its mean. In this sense, the variance is similar to the expected value in Equation 7, and its value could arise from similar cognitive mechanisms. The weighted standard deviation is given by:

$$\widehat{\sigma_{b_k}} = \sqrt{\sum_{m|\mathbf{c}_m \in \mathbf{C}} p_m \left(v_{m,k} - \widehat{\mu_{b_k}} \right)^2}. \qquad (9)$$

The beta distribution can be reparameterized in terms of $\widehat{\mu_{b_k}}$ and $\widehat{\sigma_{b_k}}$ as follows: $\alpha = \widehat{\mu_{b_k}} [x-1]$, $\beta = \left[1 - \widehat{\mu_{b_k}}\right][x-1]$, where $x = \widehat{\mu_{b_k}}(1-\widehat{\mu_{b_k}})/\sigma_{b_k}^2$.

A different retrieval request is used in the construction of the personal belief distribution and the social norm distribution (see Figure 1). For the personal belief distribution, the *source* slot has a value of *personal*. For the social norm distribution, the *source* slot has a value of *social norm*. Other slot value pairs can be included in the retrieval quest as needed. Formally, the retrieval request for distribution $z \in \{p,s\}$ can be stated generally as:

$$\mathbf{r}_z = \{(\text{source}, v_z), (s_i, v_i)\}_{i \in I_{r_z}}, \qquad (10)$$

where, as before, p corresponds to personal belief and s corresponds to social norm. Once the distributions are computed through the blending mechanism, the public belief/behavior is compute through Equations 1 and 2 as specified in SST.

Simulations

Simulation 1: Memory Dynamics and Stochasticity

The goal of Simulation 1 is twofold: (1) to show how stochasticity in belief expression arises through memory mechanisms, and (2) investigate how memory decay and experience modulate public belief distributions. Recall that SST assumes private and social norm beliefs follow a static distribution, which does not change across time or with experience. In addition, SST lacks a mechanism to produce variability in publicly expressed beliefs. However, the integrated ACT-R/SST model is sensitive to memory dynamics. Simulation 1a varied the delay between learning and public belief expression to examine the effect of decay. Simulation 1b varied the number of additional experiences with the social norm.

For both simulations, we enabled base-level learning in Equation 5 and simulated a 100 second learning phase in which the model encoded a belief every 5 seconds for a total of 10 private beliefs and 10 social norm beliefs. Private beliefs were sampled from $b_p \sim \text{beta}(4,9)$, whereas social norm

beliefs were sampled from $b_s \sim \text{beta}(6, 2.2)$. We fixed the parameters to the following values: $d = .50$, $\delta = 1$, $\eta = .20$, $w = .50$, and $\gamma = 20$. After the learning phase, we simulated the model 10,000 times under the conditions described below to obtain a distribution of public beliefs.

In Simulation 1a, the key manipulation in the simulation was the delay, Δ_t, between the learning phase and the generation of public beliefs: $\Delta_t = 10$ vs. $\Delta_t = 100$. In Simulation 1b, the key manipulation was the number of additional interactions to update the social norm: 1 vs. 10. For each interaction, a chunk for a social norm was randomly selected after a 5 second delay to emulate a social interaction (e.g., a person expressing a belief). The number of uses of the selected chunk was incremented to strengthen its activation.
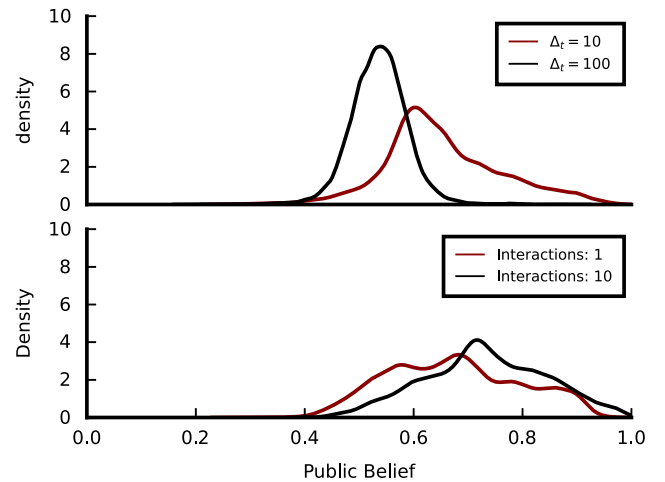
Figure 2: Top: the effect of memory decay on publicly displayed beliefs as a function of time. Δ_t denotes the time in seconds following the encoding phase. Bottom: the effect of social interactions (i.e., chunk uses) on public belief distributions.

Results There are three noteworthy findings based on Figure 2. First, in both simulations, SST-ACT-R produces distributions of utility maximizing public beliefs rather than a single, deterministic value, which is more consistent with response variability found in human behavior. Second, in Simulation 1a, after a delay of $\Delta_t = 10$, the public belief distribution was diffuse and centered near the social norm distribution, i.e. weakly conservative. After a delay of $\Delta_t = 100$, the public belief distribution changed, narrowing and shifting towards the center of the scale. The shape and dispersion of the distributions varied according to the specific private and social norm beliefs sampled during the learning phase. However, the prevailing pattern is similar to the one illustrated in Figure 2. Third, as expected, in Simulation 1b, the public belief distribution shifted towards the social norm (i.e., becoming more conservative) with increased social interactions.

Simulation 2: Social Influence Dynamics

The goal of Simulation 2 was to showcase a novel behavior emerging from the integration of SST and ACT-R—namely, social influence dynamics whereby beliefs evolve through a combination of social evaluation mechanisms and memory mechanisms. We simulated an agent-based model in which a small group of 10 interacting agents based on SST-ACT-R encoded their interactions into memory. Half of the agents were liberal leaning whereas the other half were conservative leaning.
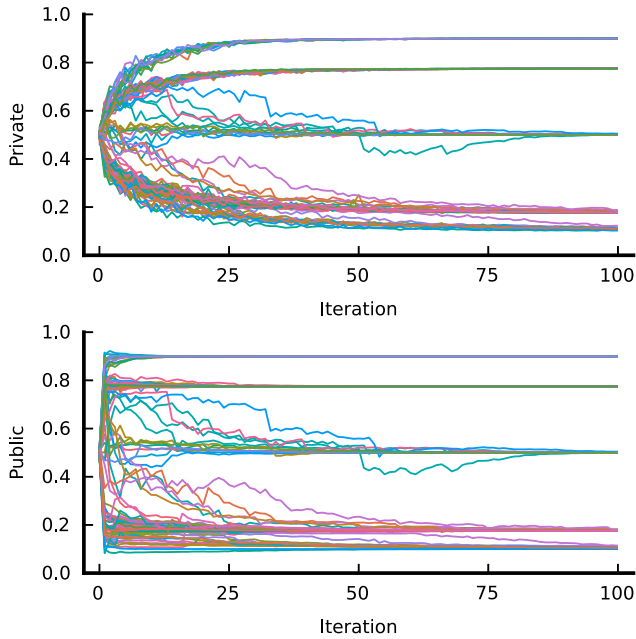
Figure 3: Social influence dynamics: the evolution of private and public beliefs averaged across agents. Each model run is denoted by a separate colored line.

Each model simulation consisted of 100 iterations and each iteration involved one step per agent. On each step, one agent was designated as the *speaker* and the other agents were designated as *listeners*. An issue was randomly selected from a set of 20, prompting the speaker to generate a public belief based using a blended retrieval over beliefs in memory with $\mathbf{r}_z = \{(\text{source}, v_z), (\text{issue}, v)\}$ serving as a retrieval request. The other agents simply encoded (i.e., listened to) the speaker's public belief. Unlike Simulation 1, the speaker encoded its publicly expressed belief with (source,public) to distinguish between public and private beliefs. By contrast, the listeners encoded the same belief with (source,social norm). After each iteration, we queried each agent's private and public beliefs and recorded the mean and standard deviation across agents.

As in Simulation 1, we fixed the parameters to the following values: $d = .50$, $\delta = 1$, $\eta = .20$, $w = .50$, and $\gamma = 20$. For values associated with the *source* slot, dissimilarity was coded as 0 for matching values, .5 for *private* vs. *public* and *public* vs. *social norm*, and 1 for *private* vs. *social norm*.

Dissimilarity for other slots was coded according to Equation 6 as before. Both liberal and conservative leaning agents were with initialized with 20 chunks which had 100 prior uses each. Private beliefs were sampled from beta(2, 8) for liberal leaning agents and beta(8, 2) for conservative leaning agents. We repeated the simulation 100 times to ensure sufficient variability in behavior.

Results Four noteworthy findings are worth unpacking. First, as shown in Figure 3, the social influence dynamics exhibit an interesting striation of both private and public beliefs into five distinct bands spanning the belief spectrum. Second, public beliefs change more rapidly than private beliefs. One contributing factor is that private beliefs are diffuse, allowing public beliefs to quickly gravitate towards an emerging social norm. By contrast, private beliefs change more slowly, in part, because they benefited from frequency effects. In addition, source confusion also contributes to the effect, as evidenced by a decrease in striation when the mismatch penalty parameter δ is increased (not shown). A third noteworthy finding is the rapid convergence onto a social norm as indicated by the decrease in the standard deviation of beliefs in Figure 4. The rate of convergence slows as the mismatch penalty parameter δ is increased, again suggesting the importance of source confusion. Finally, public and private beliefs track each other strongly, as indicated by a mean correlation of .69 (SD = .15) across simulation runs.
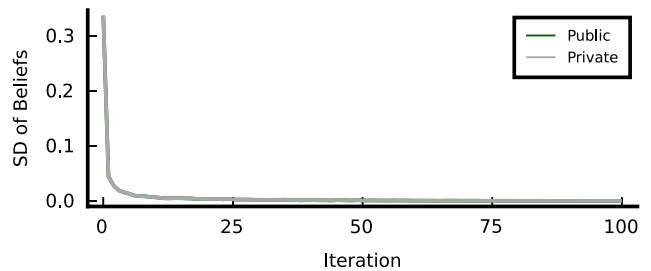
Figure 4: Social norm convergence as measured by the mean standard deviation of beliefs across runs of the simulation.

Discussion

Our research goal was to integrate SST and ACT-R to provide a detailed account of the learning and memory dynamics underlying social influence and comparison processes. The component theories of the integrated SST-ACT-R model are characterized by complementary strengths and weaknesses. On one hand, ACT-R provides a detailed account of memory and learning, but lacks a mechanism for social influence and comparison. On the other hand, the converse is true for SST: it lacks a detailed account of learning and memory, but describes the mechanisms underlying social influence and comparison. The integrated SST-ACT-R model has two noteworthy benefits: (1) it provides a more detailed account of social influence and comparison than either component theory in isolation, (2) it predicts novel dynamics in beliefs and be-

havior which can be tested in future research.

We argue that the integration of SST and ACT-R should be part of a broader research goal to extend CAs into the domain of social psychology. Integration efforts have several potential benefits. First, as we noted above, integration can lead to novel predictions and more powerful models. Second, integration efforts may provide opportunities to test the limits of CAs. Given that many CAs such as ACT-R propose that higher-order cognition arises from a set of fixed structures and cognitive capacities, social phenomena provide a challenging testbed for testing the generalizability of CAs. What makes a testbed of social phenomena particularly powerful and informative is that it is far removed from the traditional cognition (e.g., memory, perception, reasoning etc.). Thus, it is not plausible to argue that CAs have been engineered to account for social phenomena. A third benefit is that integration efforts increase the presence formal modeling in social psychology where they traditionally have less representation, leading to several benefits, such as explicating assumptions and sharpening research questions. Finally, from the perspective of public policy, the detailed mechanisms underlying integrated models have the potential to inform and improve interventions and policy prescriptions.

Limitations

The integrated SST-ACT-R model is preliminary and has several limitations worth noting. First, as stated above, one must make assumptions about functional form of the belief distributions when the number of chunks is low. Such auxiliary assumptions may introduce additional uncertainty into the model. One workaround might be to design a task in which the distributions are known, or the utility function can be estimated from large number of chunks non-parametrically.

It is worth pointing out that our preliminary exploration of the ACT-R-SST model is far from comprehensive. We plan to perform more simulation studies to understand the how various parameters modulate behavior of the model. In subsequent simulation studies, we found that increasing the mismatch penalty parameter causes public and private beliefs to evolve at a slower rate. Thus, the model can predict slower rates of belief evolution than illustrated above.

Another limitation is that the current model assumes dynamics are driven passively by memory mechanisms. However, in some cases, dynamics might also be driven by schemas, emotion, or active processes, such as motivated reasoning. Empirical tests of the model are required to establish these boundary conditions.

Finally, the integrated model introduces additional complexity in terms of increased parameters and questions about memory representation. One approach to mitigate increased complexity is to use default parameter values where possible. Using default parameter values is a valid approach for reducing complexity if they have strong support from prior research. In other cases, it might be possible to leverage these complications as opportunities to formulate novel research questions.

Conclusion

The present research builds upon a long tradition of using integrative approaches to develop productive research programs (e.g., Newell, 1990). In line with previous research, we demonstrated how integrative approaches can lead to emergent behaviors and novel predictions, which provides a basis for future empirical testing. Our hope is that our integration of SST and ACT-R serves as a stepping stone for future extensions of CAs into the domain of social psychology.

Acknowledgments

References

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2).

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

Brown, G. D., Lewandowsky, S., & Huang, Z. (2022). Social sampling and expressed attitudes: Authenticity preference and social extremeness aversion lead to social norm effects and polarization. *Psychological Review*, *129*(1), 18–48.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., . . . Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29.

Fridman, N., & Kaminka, G. A. (2011). Towards a computational model of social comparison: Some implications for the cognitive architecture. *Cognitive Systems Research*, *12*(2), 186–197.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523–551.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Kuran, T. (1997). *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior

theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Nyamsuren, E., & Taatgen, N. A. (2013). Pre-attentive and attentive vision module. *Cognitive Systems Research*, *24*, 62–71. doi: 10.1016/j.cogsys.2012.12.010

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301.

Stevens, C. A., Taatgen, N. A., & Cnossen, F. (2016). Instance-based models of metacognition in the prisoner's dilemma. *Topics in Cognitive Science*, *8*(1), 322–334.

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: systematic review. *Journal of Medical Internet Research*, *23*(1), e17187.

Sun, R. (2007). The importance of cognitive architectures: An analysis based on CLARION. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*(2), 159–193.

Hey Pentti, We Did It!: A Fully Vector-Symbolic Lisp

Eilene Tomkins-Flanagan (eilenetomkinsflanaga@cmail.carleton.ca)
Mary Alexandria Kelly (mary.kelly4@carleton.ca)
Department of Cognitive Science, Carleton University
1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada

Abstract

Kanerva (2014) suggested that it would be possible to construct a complete Lisp out of a vector-symbolic architecture. We present the general form of a vector-symbolic representation of the five Lisp elementary functions, lambda expressions, and other auxiliary functions, found in the Lisp 1.5 specification (McCarthy, 1960), which is near minimal and sufficient for Turing-completeness. Our specific implementation uses holographic reduced representations (Plate, 1995), with a modern Hopfield network (Ramsauer et al., 2020) employed as a cleanup memory. Lisp, as all Turing-complete languages, is a Cartesian closed category (nLab authors, 2024), unusual in its proximity to the mathematical abstraction. We discuss the mathematics, the purpose, and the significance of demonstrating vector-symbolic architectures' Cartesian-closedness, as well as the importance of explicitly including cleanup memories in the specification of the architecture.

Keywords: vector-symbolic architecture; Lisp; holographic reduced representations; cartesian closed category; modern hopfield network

At Clojure/Conj 2023, the conference of the Clojure programming language, Meier (2023) introduced vector-symbolic architectures to the Clojure community. Her presentation echoed a motif oft heard listening to programmers who discover vector-symbolic architectures (VSAs) for the first time; namely, VSAs' unusual properties and computational niceties are objects of fascination, but it is not immediately obvious what good a vector-symbolic architecture does for the programmer. We present an existence proof that VSAs are completely general computational tools. In technical terms, VSAs can do anything one wants. In pragmatic terms, "technically anything" does not answer questions of naturalness and ease of representation. To answer practical questions, VSAs have been used most frequently in representing human cognition, fruitfully in simultaneous localization and mapping (SLAM), and, pertinent to our analysis, promisingly in the syntactic manipulation of neural network states.

In an remark at the end of her talk, Meier mentioned a "challenge", issued by Kanerva, in "one of his papers", to implement Lisp using exclusively a vector-symbolic architecture representing all the language's expressions. However, the exact words read in the talk as a challenge, "One could create a 'High dimensional computing-Lisp'", do not seem to have been written by Kanerva. This apparent mistake is not Meier's, as those exact words have been published in Neubert, Schubert, and Protzel (2019), who attribute the enclosed quote to Kanerva (2014). While the quote does not

appear in Kanerva's paper, the mistake is plausibly a case of miscitation of something said during the associated conference talk, and in any case it is not serious. Kanerva's paper, disappointingly, does not include any challenge, but rather a discussion of how a vector-symbolic Lisp might be implemented, coupled with a loose specification of some of the tools that might be required to do so. We are going to pretend that counts as a challenge, and fully specify a Lisp language in terms of a generic VSA[1].

Hold Up, What's a VSA?

A vector-symbolic architecture is an algebra (i.e., a vector space with a bilinear product),

1. that is closed under the product $\otimes : V \times V \to V$ (i.e., if $u \otimes v = w$, then $u, v, w \in V$)
2. whose product has an "approximate inverse" $\overline{\otimes}$ that, given a product w and one of its operands u or v, yields a vector correlated with the other operand
3. for which there is a dogma for selecting vectors from the space to be treated as atomic "symbols",
4. that is paired with a memory system $\mathcal{M}$ that stores an inventory of known symbols for retrieval after lossy operations (e.g., inversion), that can be recalled from $\mathcal{M}(p)$, and which is appendable $\mathcal{M} \leftarrow t$, and
5. possesses a measure of the correlation (a.k.a., similarity) of two vectors, $\mathbf{sim}(u, v) \in [-1, 1]$, where 1 and -1 imply that u, v are colinear, 0 that they are linearly independent.

The $+$ and $\otimes$ operators behave analogously to disjunction and conjunction, or set-theoretic union and intersection. Additionally, VSAs may have an analogue for negation, often the vector rejection on Euclidean space $\mathbf{rej}_v(u)$ (Widdows & Cohen, 2014), and permutations Π, which are typically used to introduce asymmetry to the product operator, by applying different permutations to the operands. For a detailed review of vector symbolic architectures, see Kleyko, Rachkovskij, Osipov, and Rahimi (2022, 2023). Heddes et al. (2023) develop a software library for applying VSAs based on Torch.

In our implementation, we use holographic reduced representations (Plate, 1995). They are defined over Euclidean space $\mathbb{R}^n$, and have circular convolution as their product, co-

---

[1]Our implementation may be found at `https://github.com/eilene-ftf/holis`

sine as their similarity, and atomic symbols sampled from a Gaussian distribution. Our memory system is a lookup table.

Some VSAs (e.g., Kanerva, 1996) are not defined over vector spaces per se, or otherwise relax some of the above properties, but behave sufficiently similarly to be used in a similar way. The programmer's choice of VSA comes down to preference and different computational conveniences. For the most part, all VSAs are as good as all others.

Vector-symbolic architectures are an answer to an old tension in cognitive science between the actual machinery of the brain and properties cognition is believed to necessarily possess. Brain states, to the one side, are described in terms of the activity of multiple cell populations, and they exist over a fixed number of cells. Information is typically assumed to be distributed over measured populations, degrading gracefully and uniformly when cells are disabled at random. To the other, Fodor and Pylyshyn (1988) made a compelling case that central cognition must have states that behave like discrete symbols, that can be strung together in a combinatorial syntax. But, the traditional tool for representing such syntaxes, computer memory, uses strings of bits for individual symbols, composed by concatenation into ever-longer strings.

With disjunction (sum), conjunction (product), inverse, and similarity operations, plus a cleanup memory, a VSA is sufficient to describe any syntax one could want, to a finite precision. Thus, VSAs appear to satisfy the cogntive scientist's parallel demands for syntax and biomimicry.

Cartejian Clojure and Lisp

A Cartesian closed category (nLab authors, 2024) is the mathematical generalization of what it means to compute. It generalizes the equivalence of universal Turing machines (Li & Vitányi, 2008, ch. 1) with other definitions of computation, set theory, first-order logic, and, of interest to us, the recursively enumerable languages **RE** (Chomsky, 1955). All instances of a Cartesian closed category have the preceding equivalences; to say that C is Cartesian closed is also to say that it is Turing-complete. It follows that C can define **RE**, and, therefore, any syntax, as the language generated by any syntax rules, or grammar, is a subset of that generated by **RE**.

Categories have two contents: *objects* and *morphisms*[2]. For example, while we normally treat vector spaces as sets of vectors augmented with some functions, they are equally categories that *include* both vectors (their objects) and functions (their morphisms). Cartesian closed categories in particular are useful because they are very simple, and so it is usually easy to demonstrate that a formal system is Cartesian closed.

In a Cartesian closed category C, there is (1) one object, called a terminal object **1** (so-named because there is a morphism from every object in the category to it). There is also (2) a product that can compose any two objects, under which C is closed, i.e., if A, B are objects in C, the product of any

objects $A \times B$ in C is also an object in C. (3) The functions $A \to B$ on objects in C are together an object in C, written B^A. (4) A morphism that evaluates functions in C, parameterized by objects in C, **eval**$_C : B^A \times A \to B$, is in C.

These four properties give us four tests for whether some formal system S is Cartesian closed. S must have at least one base data object, and we should be able to transform any expression into it (1). S must have some means to compose arbitrary expressions from its objects, that are data objects still usable by S (2). S must be able to express functions that may map any objects to any others, and those functions must be representable as data objects (3). Finally, we should be able to describe a complete interpreter for S, in terms of S(4).

Keen readers will have noticed why the above defines computation. We have some base symbols; we may construct sequences of symbols, any length; we can specify any function that transforms sequences to other sequences; and, we can evaluate those functions. That is pretty much a description of a universal Turing machine (see Li and Vitányi, ch. 1).

McCarthy (1960) described Lisp for the 1.5th time, giving us the mother document of all subsequent Lisp dialects. Its simplicity will enable us to complete our "challenge" without taking up a whole book. To demonstrate that VSAs can compute, we need only implement the five "elementary" functions of Lisp 1.5, plus some other functions that can be recursively defined in terms of the others. The elementary functions are: CONS, CAR, CDR, EQ, and ATOM. Additionally, there are LAMBDA, COND, and LABEL. LAMBDA is the most important, as it allows us to define lambda expressions, i.e., arbitrary functions.

In Lisp 1.5, a tuple is represented in the form (a . b), where a and b are either atomic symbols (written as alphanumeric sequences) or other tuples. A list (a b c) is equivalent to the tuple (a . (b . (c . NIL))), where NIL is an atomic symbol that represents the end of a list. Naturally, the singleton (a) is the tuple (a . NIL) and the empty list () is just NIL. The atomic symbols NIL, T, and F are always defined. We'll define the elementary functions, where lowercase letters are variables that may be any valid Lisp expression:

```
(CONS a b) = (a . b)
(CAR (a . b)) = a
(CDR (a . b)) = b
(EQ a a) = T
(EQ a b) = F
(ATOM (a . b)) = F
(ATOM a) = T
```

The preceding definitions use a pattern-matching format, such that the earlier definition takes precedence. Where the same letter is used for two variables, the variables must be identical. In plain English, CONS takes any two expressions, and constructs a tuple containing them. CAR takes a constructed pair, and yields the left element, while CDR does so with the right. EQ tests whether two atomic symbols are identical, and is undefined for non-atomic symbols. ATOM tests whether an expression is atomic. This already seems like very

---

[2]Generally, a morphism is any way objects can be related such that, if you have two morphisms f, g, you can construct $h = f \circ g$ such that $h(a) = f(g(a))$ for some appropriate notion of equivalence.

little, but armed with an understanding of Cartesian closure, it can be understood that we don't even need all of Lisp to have a Turing-complete language. We just need `CONS` (our product), `ATOM` (in case it is not immediately clear, there is a morphism from any expression `e` to `T` given by `(ATOM (ATOM e))`), and `LAMBDA` (for functions; the Lisp interpreter evaluates expressions and can be described as a Lisp expression). What makes Lisp remarkable as a point of reference is that there is almost no fat on top of the basic building blocks of a bicartesian closed category; we can describe anything we would like recursively in terms of the basic functions. To wit:

```
((LAMBDA NIL e)) = e
((LAMBDA x NIL) a) = NIL
((LAMBDA x ((CAR x) . e)) a) = (LAMBDA (CDR x)
    (a . ((LAMBDA x e) a)))
((LAMBDA x ((c . d) . e)) a) = (LAMBDA (CDR x) (
    ((LAMBDA x (c . d)) a) . ((LAMBDA x e) a)))
((LAMBDA x e) a) = (LAMBDA (CDR x) ((CAR e) .
    ((LAMBDA x (CDR e)) a)))
```

The preceding recursively defines lambda expressions entirely in terms of the Lisp elementary functions, provided that arguments are always curried[3]. The above recursive definition has five cases, where any time `LAMBDA` is called, the earliest definition that fits the arguments takes precedence. A lambda expression is a three element list, containing `LAMBDA`, a list of parameters `x`, and an enclosed expression `e`. At base, `(LAMBDA x e)` does nothing, but it can be called on one argument `a`, which may be `NIL`, `((LAMBDA x e) a)`, and then it is evaluated over `a`, returning either a lambda expression where `a` is substituted for all instances of the first parameter, or, if there are no arguments left, the resultant body expression with all substitutions made. In our definition, all lambda expressions are always curried, so a function on three arguments `a`, `b`, `c` is called as `(((((LAMBDA x e) a) b) c))`, with the final call being implicitly on the single argument `NIL`, as `NIL` terminates all lists. The parameters, `x`, are a list that is assumed to consist of atomic symbols. `LAMBDA` is undefined where elements of `x` are nonatomic or duplicate.

`COND` implements conditional expressions:

```
(COND ((T . q) e)) = q
(COND e) = (COND (CDR e))
```

The way conditionals work is pretty straightforward. We write some implications, and when evaluated, we take the first branch whose condition is satisfied after evaluation.

It is worth noting that this form of recursive definition is useful for its terseness, but it is not proper to LISP 1.5, which would require the use of a `DEFINE` pseudo-function to instantiate a function definition. `DEFINE` is not one of the elementary functions because it just maps a name to an expression in system memory. Recursive expressions are possible *without*

---

[3]A more Lisp-appropriate definition might have been written such that arguments do not have to be curried, but this version was chosen for ease of presentation.

using `DEFINE`, so the above effects can be achieved (if not persistently named) by using the `LABEL` function. `LABEL` is the fundamental tool by which recursion is achieved in the Lisp 1.5 specification, but we choose to omit it due to redundancy.

What we notice in the Lisp 1.5 specification is that there is remarkable inclarity as to what is core to the language and how things are formally defined. Understanding Cartesian closed categories, however, helps to clear up some details. We have chosen to define the parts of Lisp that are elementary, plus lambda expressions, and functions that make programming minimally less painful: `QUOTE`, `COND`, and `DEFINE`. We are now prepared to describe the Lisp VSA.

The Lisp VSA

The logic of a LISP VSA is straightforward. We are going to map all the elementary functions of Lisp to operations in a vector-symbolic architecture. This proves remarkably straightforward. An interpreter for the Lisp VSA reads a Lisp program and, instead of executing, e.g., `CONS`, over two bytes in order to make a two-byte array in memory, it will apply the vector-symbolic `CONS` over two vectors in order to create a joint representation of the pair, as a single vector.

One detail that has not been addressed is how atomic symbols are to be constructed. As the dimension of a space grows, fixing one vector and choosing another vector arbitrarily, the expected value of their similarity goes to 0, and vectors with nonnegligible similarity are exceedingly rare (Kainen & Kůrková, 1993). In holographic reduced representations (Plate, 1995), we sample vectors on $\mathbb{R}^n$ from a normal distribution, with $\mu = 0$ and $\sigma = \frac{1}{\sqrt{n}}$, producing vectors v with $\mathbb{E}[||v||] = 1$. Thus, we can have many more base symbols than dimensions in the space, all *nearly* linearly independent[4], which would not be the case if they were truly orthogonal. For Euclidean vectors, it typical to see $n \in \{2^k, k \in [6, 12]\}$.

Kanerva (2014) suggested representing lists by permuting one operand then adding the two operands together. By keeping a fixed permutation in memory, the united representation is most similar to its unpermuted operand by default, and then, by applying the inverse permutation, winds up most similar to its permuted operand, with fixed permutation Π.

$$\mathbf{cons}(a, b) = a + \Pi(b)$$
$$\mathbf{car}(c) = \mathcal{M}(c)$$
$$\mathbf{cdr}(c) = \mathcal{M}(\Pi^{-1}(c))$$

This method has some flaws if we care about retrieval, however. Taking advantage of the property just described, either operand can be retrieved from a tuple very simply. But, problematically, if one wishes to make a list of arbitrarily many elements, one needs to store sublists in memory. Once

---

[4]We can calculate the expected variance of pairwise $\mathbf{sim}(a, b)$ for an arbitrary overcomplete basis B (i.e., a finite sample of $\mathbb{R}^n$ where $|B| > n$) with $a \neq b \in B \subset \mathbb{R}^n$ exactly, but that calculation is outside the scope of this paper. A commonly used "margin of safety" expects $\mathbf{sim}(a, b) \in (-0.2, 0.2)$, but for $n \geq 512$ the actual expected variance is much smaller, even for large $|B|$.

one stores a list in memory, however, the vector to which that list is most similar is itself. It becomes necessary to do something to tag *both* operands, such that each operand and their disjunction may be reliably distinguished in memory.

Permutation is also a redundant operation if we are implementing a Lisp. Although it is often used to make the conjunction operator $\otimes$ asymmetric, this behaviour is not necessary if we are implementing a Cartesian closed category, as the role of taking an operation that builds a joint representation of two operands, and making it asymmetrical, is already satisfied in the VSA algebra. In the Lisp VSA, the product's job is making a combined representation of two objects that is dissimilar to either, but both remain retrievable, using their pair and a cleanup memory, which is exactly what permutation is doing in Kanerva. As such, we will leave out the permutation operator and work just with $+, \otimes, \overline{\otimes}, \mathbf{sim}, \mathcal{M}$,

We define additional operators for convenience. $\mathcal{F}$ is another cleanup memory, that separates global function bindings from the inventory of retrievable expressions. $\nu(v) = \frac{v}{||v||}$ divides v by its magnitude, such that $\mathbf{sim}(\nu(v), v) = 1$, $||\nu(v)|| = 1$. $\oplus$ is a variant addition operator that saturates on both an upper and lower threshold, respectively $\theta_\uparrow, \theta_\downarrow$:

$$a \oplus b = (\mathbb{E}[||a||] > \theta_\uparrow)a$$
$$+ (\mathbb{E}[||a||] < \theta_\downarrow)b \qquad (1)$$
$$+ (\theta_\downarrow \leq \mathbb{E}[||a||] \leq \theta_\uparrow) \, \nu(a+b)$$

Because $\oplus$ is defined using three mutually exclusive cases, the operands a and b can be lazily evaluated, such that the values are only computed if they are needed. Defined assuming lazy evaluation, $\oplus$ is a useful operator for writing recursive definitions. Setting $\theta_\uparrow$ and $\theta_\downarrow$ respectively a little under 1 and a little over 0, and the expected magnitude of all vectors is 1, expressions written with $\oplus$ are meant to be read as evaluating the left operand if a test multiplying it succeeds (the test yields a scalar value $\alpha = \mathbb{E}[||a||] > \theta_\uparrow$), and evaluating the right operand if it fails ($\alpha < \theta_\downarrow$). Several additional atomic expressions are used in the preceding definitions, notably L and R, which mark the left hand and right hand sides of a tuple, as well as φ, which marks that a vector is nonatomic.

$\mathbf{f}(a)$ marks a call to programmer-defined function $\mathbf{f}$, and requires some special treatment, as $\mathbf{cons}(\mathbf{f}, a)$ is equivalent in our notation to $\mathbf{f}(a)$. What the notation means is that the interpreter should leave the list including $\mathbf{f}$ as an unevaluated expression if $\mathbf{f}$ is not in the function namespace.

Below are the definitions of the Lisp VSA. Single unbolded lowercase letters refer to variables that may contain arbitrary Lisp expressions, but typically they are expected to be of a certain form and lead to undefined behaviour when not of that form. Bolded words and letters refer to function names, and are expected to always be atomic. Functions in general are called by simply using their name, and so all function calls are marked by atomic symbols at the head of a list of arguments, except in the case of lambda expressions, which are three-element lists. Lisp expressions of the form (F a b ...) are

translated to our notation as $\mathbf{f}(a, b, ..., t)$, where the last element t is always the tail of the list of arguments. Recalling that lists are recursively nested tuples, (F a b ...) is equivalent to (F . (a . (b ...))), and likewise $\mathbf{f}(a)(b)... = \mathbf{f}(\mathbf{cons}(a, \mathbf{cons}(b, ...))) = \mathbf{cons}(\mathbf{f}, (\mathbf{cons}(a, \mathbf{cons}(b, ...))))$ in our notation. Programmer-defined functions are always fully curried. The special case of (F) is, following the definition of lists, equivalent to $\mathbf{f}(\text{NIL})$. Therefore, we never technically have functions on no arguments. Here are the definitions:

$$\mathbf{cons}(a, b, \_) := \nu(L \otimes a + R \otimes b + \varphi) \mid \mathcal{M} \leftleftarrows a, b \qquad (2)$$

$$\mathbf{car}(a) := \mathcal{M}(L \overline{\otimes} a) \qquad (3)$$

$$\mathbf{cdr}(a) := \mathcal{M}(R \overline{\otimes} a) \qquad (4)$$

$$\mathbf{eq}(a, b, \_) := \mathbf{sim}(a, b)T + (1 - \mathbf{sim}(a, b))F \qquad (5)$$

$$\mathbf{atom}(a, n) := \mathbf{sim}(n, \text{NIL})\mathcal{M}(\mathbf{sim}(a, \varphi)F$$
$$+ (2\theta_\downarrow - \mathbf{sim}(a, \varphi))^+ T) \qquad (6)$$
$$+ (2\theta_\downarrow - \mathbf{sim}(n, \text{NIL}))^+ F$$

$$\mathbf{define}(a, e, \_) := * \mid \mathcal{F} \leftleftarrows \mathbf{cons}(a, e) \qquad (7)$$

$$\mathbf{cond}(r) := \mathbf{sim}(\mathbf{car}(\mathbf{car}(r)), T)\mathbf{cdr}(\mathbf{car}(r))$$
$$\oplus \mathbf{cond}(\mathbf{cdr}(r)) \qquad (8)$$

$$(\lambda(x, e))(a) := \mathbf{sim}(x, \text{NIL})e$$
$$\oplus \mathbf{sim}(e, \text{NIL})\text{NIL} \qquad (9)$$
$$\oplus \lambda(\mathbf{cdr}(x), \lambda \mathbf{s}(x, e)(a))$$

$$(\lambda \mathbf{s}(x, e))(a) := \mathbf{sim}(x, \text{NIL})e$$
$$\oplus \mathbf{sim}(e, \text{NIL})\text{NIL}$$
$$\oplus \mathbf{sim}(\mathbf{car}(x), e)\mathbf{car}(a)$$
$$\oplus \mathbf{sim}(\mathbf{atom}(e), T)e$$
$$\oplus \mathbf{sim}(\mathbf{car}(x), \mathbf{car}(e))$$
$$\quad \mathbf{cons}(\mathbf{car}(a), \lambda \mathbf{s}(x, \mathbf{cdr}(e))(a))$$
$$\oplus \mathbf{sim}(\mathbf{atom}(\mathbf{car}(e)), F) \qquad (10)$$
$$\quad \mathbf{cons}($$
$$\quad\quad (\lambda \mathbf{s}(x, \mathbf{car}(e)))(a),$$
$$\quad\quad (\lambda \mathbf{s}(x, \mathbf{cdr}(e)))(a)$$
$$\quad)$$
$$\oplus \mathbf{cons}(\mathbf{car}(e), (\lambda \mathbf{s}(x, \mathbf{cdr}(e)))(a))$$

$$\mathbf{f}(a) := \mathbf{sim}(\mathbf{f}, \mathbf{car}(\mathcal{F}(L \otimes \mathbf{f})))$$
$$\quad \mathbf{cons}(\mathbf{cdr}(\mathcal{F}(L \otimes \mathbf{f})), a) \qquad (11)$$
$$\oplus \mathbf{cons}(\mathbf{f}, a)$$

With some minor modifications due to simplifications of the specification, the above definitions can be used to implement the Lisp 1.5 interpreter (McCarthy, Appendix B).

What is particularly notable in the above definitions is the frequency and fundamentalness of the use of cleanup memories. Every VSA has a cleanup memory, but usually, the cleanup memory relies on a big matrix M that

stores every known symbol as an approximately unit length row vector. Thus, on $\mathbb{R}^n$, the cleanup memory is $\mathcal{M}(p) = \mathbf{argmax}(pM^T)M$, where p is a probe vector to be "cleaned up", by retrieving its nearest neighbour from memory. Because of the historic difficulty of implementing both time- and space-efficient cleanup memories, and a low appraisal of the biological significance of "memory is a big lookup table and you test every entry in order to retrieve the one you want", the choice of cleanup memory being used by any given VSA is an embarrassment one typically glosses over (e.g. while Kanerva, 2014 discusses cleanup memories, they are not explicit in his algebraic notation, and are left as a black box in his diagrams). We emphasize the explicit notation of cleanup memories, because they are essential for achieving features of VSAs in frequent day-to-day use, because different cleanup memories have distinctive computational properties that may fit some applications better than others, and because memories with certain computational properties are essential to achieving Turing-completeness in our Lisp.

Kanerva (2014), following Eliasmith et al. (2012), refers to the vectors of a VSA as "pointers". That is because, partly, a probe in the memory can be taken to "reference" its nearest neighbour; a trace can be looked up using any of the points in space near it. Traditional memory pointers similarly "probe" memory, though, in general, what is at the probed memory location need not have high mutual information with the probe. Variant cleanup memories can also be defined that are similarly *heteroassociative*, making probes much more like traditional pointers. One glaring flaw appears in the pointer analogy, however: Where d is the number of stored traces, retrieval from computer memory is $O(\log(d))$, if d is close to the total available space ($O(1)$ if significantly less). Probing a traditional cleanup memory is at least $O(d)$. In fact, because probing memory *requires* traversing all stored traces to test for similarity, the lookup is also at least $\Omega(d)$. It is not an issue of principle versus practice either; because VSAs use vectors of extremely high dimensionality, comparisons take a long time, and because one is often storing thousands or millions of vectors in memory for practical applications, one is really getting one's n's worth of comparisons in.

So What is to be Done?

Neubert et al. make a second apparent misattribution to Kanerva (2014), which is also fruitful to pretend was written as attributed. They suggest the possibility of another type of cleanup memory: an attractor neural network. In such networks, information is often (though not always) distributed over the network's weights, which makes them robust to noise or damage, as the vector representations of VSAs are robust. Attractor networks feature interacting cells converging to stable patterns over time, a tantalizingly brain-like property. However, most attractor networks in use are no more time or memory efficient than a big matrix. The Hopfield network (Hopfield, 1982, made continuous in Hopfield, 1984) has a storage capacity of $O(n)$ with respect to its input di-

mensions. Hopfield networks work almost *exactly like* the big matrix format, with a different *activation function* (above, our activation function was **argmax**) and the proviso that the network's outputs may be fed back into it, until it converges[5].

Another appealing option is to use the match networks of Grossberg (2021), as they've seen some success in modelling human brains, and also claim to solve retroactive interference. Unfortunately, they also look like the big matrix approach of before[6], and they eliminate retroactive interference by "gating" gradient descent, with a function that updates only on one row at a time, prohibiting the storage of more than $O(n)$ traces or retrieving them in less than $\Omega(d)$ time, if traces are assumed to have low mutual information.

If we relax the requirement that traces be near-orthogonal, better results may be obtained. Ororbia and Kelly (2023) use a continuous variant of MINERVA2 (Hintzman, 1984) as the memory system of a reinforcement learning agent. MINERVA2 also resembles the "big matrix" memory: $\mathcal{M}(p) = (pM^T)^\rho M$ where ρ is an odd integer power. ρ can be allowed to be a real number using the variant equation $\mathcal{M}(p) = \mathbf{sgn}(\xi)(\mathbf{sgn}(\xi)\xi)^\rho M$ where $\xi = pM^T$. Traces are still inserted row-wise, but Ororbia and Kelly do not expect to retrieve traces exactly as-stored, and rather interpolate between stored traces using probes similar to several of them. They also employ a forgetting mechanism: when information is unused, it fades out of memory. Thus, the size of memory is capped, without running out of space. Their system is not strictly vector-symbolic; there is no syntactic manipulation. But, if atomic symbols are allowed to be correlated and we permit forgetting, similarly advantageous properties may be usable. One reason to specify the exact cleanup memory used in one's VSA is that its space, timing, and information loss characteristics are very relevant topics for study. Different tradeoffs of characteristics might significantly affect the behaviour of the VSA in a specific use-case.

For our vector-symbolic Lisp, MINERVA2 is inadequate, at least without significantly modifying the specification. Let us reflect on the general form of the cleanup memories we have looked at: $\mathcal{M}(p) = \tau(\sigma(\beta pM^T)M + q)$ with activation function σ, normalization function τ, scalar constant β, and some added factor q. In most of the preceding cases, τ has been the identity, $\beta = 1$ and $q = 0$. M is an $m \times n$ matrix, where m is typically $O(d)$, each row storing one n-dimensional trace. Ideally, we want M to distribute information about retrievable traces, as in the case of Hopfield networks and MINERVA2; we want to retrieve traces exactly as-stored, as in the big matrix case; we want to store a number of traces that is superlinear relative to the input dimension n, both for the sake of having a cleanup memory with a great capacity, and for improving the memory's timing characteristics. Capacity is important, because our Lisp relies so heavily

---

[5]Hopfield called for updating neuron activations at random, but both bulk and random updates converge to the same outcomes.

[6]Converging to $\mathcal{M}(p) = \tau(\sigma(pM^T)M + p)$, where σ behaves similarly to **argmax**, and $\tau(v) = \varsigma(\frac{v}{||v||_1})$ with logistic function ς.

on memory to allow for a program to be written with arbitrary functions, and because the set of functions on a *S* is the powerset $\mathcal{P}(S)$; if arbitrary programs are allowed, we need to be able to store and retrieve arbitrary sequence-to-sequence maps, requiring a memory system that is at least exponential in storage capacity with respect to the input dimension. Just one candidate cleanup memory fits the bill: the modern Hopfield Network (MHN) (Ramsauer et al., 2020).

Mathematically demonstrating an exponential storage capacity and therefore $O(\log(d))$ lookup time, Ramsauer et al. describe a neural network with a familiar form:

$$\mathcal{M}(p) = \textbf{softmax}(\beta \, pM^T)M$$

Information is distributed because stored traces are encoded in *M* using gradient descent (no gating). Special cases are the big matrix variant (equal to the limit of the MHN equation as $\beta \to \infty$) and a linear memory system (with $\beta = 0$). If β is allowed to be a function of ξ, then MINERVA2 is also a special case (after normalization) with $\beta = \frac{\rho \log(x)}{x}$ (demonstrable algebraically), but this is not a very useful example. The special cases serve to indicate that the capacity of the network is sensitive to the choice of β. Given an arbitrary program, then, some amount of tuning is necessary to make the MHN store what needs to be stored.

However, MHNs obtain an exponential capacity by way of encoding by gradient descent, so, while capacity is large and retrievals are $O(1)$, encoding is $O(n)$, and requires a backup of all stored traces, if retroactive interference is to be avoided. As it is possible for many traces to be stored at runtime in the VSA Lisp, MHNs present significant drawbacks for us, so long as they depend on a gradient descent learning rule.

And Whatfor This Lisp?

The behaviour of vector-symbolic architectures is very sensitive to the choice of cleanup memory. While memory characteristics are not typically used to demonstrate Turing-completeness, Turing-completeness makes sufficiently great demands of memory that reasonable performance requires specific memory characteristics. As the computing applications of VSAs expand, studying these characteristics and making good tradeoffs will be very important.

One application that stands out is suggested by Tamkin, Taufeeque, and Goodman (2023), who trained a transformer network to exhibit states that were decomposable into "monosemantic" vectors *S*. The semantic content of the network's output was manipulated by adding or subtracting features drawn from *S*. As such, transformer states may be made to behave as additive compositions of atomic vector symbols, of the sort syntactically composable by VSAs.

Creating a vector-symbolic Lisp has been alluded to a few times, in particular by Kanerva (2014), Smolensky (1990), and Legendre, Miyata, and Smolensky (1990). The appeal is obvious to cognitive scientists and explicit in Smolensky: We think that brain states have syntax, and we know information is distributed over them. VSAs are a means to express syntax

in terms that may describe brain states, and Lisp instantiates the most general class of syntaxes. Respecting actual neural networks, Chen et al. (2020) and Smolensky, McCoy, Fernandez, Goldrick, and Gao (2022) have put syntactic manipulation of network states into practice, with promising results.

Traditional cognitive architectures, such as ACT-R, describe memory states syntactically (Stewart & West, 2006), and take actions according to rules that are sensitive to syntactic features. Such states have already been described in terms of a VSA (Kelly, Arora, West, & Reitter, 2020), and, taking into account their rules and memory systems, these cognitive architectures are already Turing-complete. However, it is uncommon for these cognitive architectures to treat memory states as arbitrary programs, and attempt to evaluate them. It is notable that these architectures directly descend from an attempt to describe artificial general intelligence (Newell, 1980), that the only formal description of artificial general intelligence (Hutter, 2000) expects states of memory to be arbitrary programs, that recent research obliquely referencing the latter suggests it is at least worth considering that states of human memory are such arbitrary programs (Dehaene, Al Roumi, Lakretz, Planton, & Sablé-Meyer, 2022), and that none of the preceding should be at all surprising, since Turing's universal machine was originally a description of the sorts of things people do in their head, manipulating either their memory or a piece of paper (Turing, 1950). Therefore, it may be necessary to increase the expressivity of memory states in order for existing paradigms to capture some complex human behaviour. To that end, it is only necessary to describe rules that treat certain states of memory as lambda expressions and evaluate them (as we did above). Then, memory may encode arbitrary programs, although, based on the behaviour of our own interpreter (see footnote 1), our simple design is likely not the most efficient that can be achieved.

But what is most striking is that, in several leading functional theories of consciousness, a necessary feature is one's ability to pursue one's goals by reading and manipulating of one's own internal state (Butlin et al., 2023, p. 5, Table 1, properties RPT-1, 2, HOT-2, 3, AST-1, and AE-1). By identifying atomic states, and training a network to represent compositions of states, an arbitrary syntax can be defined over the network's state space, though not all networks can be trained to make all syntaxes useful. If any useful syntaxes are possible, which it seems they are, then the states of some networks can be decomposed into sequences, so sequence-to-sequence models may become capable of reading and writing the very states that control their behaviour. Because VSAs are provably Turing-complete, there are no limits to how they can subject network states to syntactic manipulation, if those syntaxes can be encoded and learned over. If any of the noted theories surveyed in Butlin et al. are correct in requiring such auto-manipulation, *vector-symbolic architectures might even be the gateway to machine consciousness.*

References

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness.* doi: 10.48550/arXiv.2308.08708

Chen, K., Huang, Q., Palangi, H., Smolensky, P., Forbus, K., & Gao, J. (2020, 13–18 Jul). Mapping natural-language problems to formal-language solutions using structured neural representations. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 1566–1575). PMLR. Retrieved from `https://proceedings.mlr.press/v119/chen20g.html`

Chomsky, N. (1955). *The logical structure of linguistic theory*. Unpublished doctoral dissertation, University of Pennsylvania. (Published as a monograph by Plenum Press, New York, in 1975)

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*(9), 751-766. doi: 10.1016/j.tics.2022.06.010

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202-1205. doi: 10.1126/science.1225266

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1), 3-71. doi: 10.1016/0010-0277(88)90031-5

Grossberg, S. (2021). *Conscious mind, resonant brain: How each brain makes a mind*. Oxford University Press.

Heddes, M., Nunes, I., Vergés, P., Kleyko, D., Abraham, D., Givargis, T., ... Veidenbaum, A. (2023). Torchhd: An open source python library to support research on hyperdimensional computing and vector symbolic architectures. *Journal of Machine Learning Research*, *24*(255), 1–10. Retrieved from `http://jmlr.org/papers/v24/23-0300.html`

Hintzman, D. L. (1984, March). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101. doi: 10.3758/BF03202365

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554-2558. doi: 10.1073/pnas.79.8.2554

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, *81*(10), 3088-3092. doi: 10.1073/pnas.81.10.3088

Hutter, M. (2000). *Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decision theory.*

Kainen, P. C., & Kůrková, V. (1993). Quasiorthogonal dimension of euclidean spaces. *Applied Mathematics Letters*, *6*(3), 7-10. doi: 10.1016/0893-9659(93)90023-G

Kanerva, P. (1996). Binary spatter-coding of ordered K-tuples. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, & B. Sendhoff (Eds.), *Artificial neural networks — ICANN 96* (pp. 869–873). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/3-540-61510-5_146

Kanerva, P. (2014). Computing with 10,000-bit words. In *2014 52nd annual Allerton conference on communication, control, and computing (Allerton)* (p. 304-310). doi: 10.1109/ALLERTON.2014.7028470

Kelly, M. A., Arora, N., West, R. L., & Reitter, D. (2020). Holographic declarative memory: Distributional semantics as the architecture of memory. *Cognitive Science*, *44*(11), e12904. doi: 10.1111/cogs.12904

Kleyko, D., Rachkovskij, D., Osipov, E., & Rahimi, A. (2023, jan). A survey on hyperdimensional computing aka vector symbolic architectures, part ii: Applications, cognitive models, and challenges. *ACM Comput. Surv.*, *55*(9). doi: 10.1145/3558000

Kleyko, D., Rachkovskij, D. A., Osipov, E., & Rahimi, A. (2022, dec). A survey on hyperdimensional computing aka vector symbolic architectures, part i: Models and data transformations. *ACM Comput. Surv.*, *55*(6). doi: 10.1145/3538531

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Distributed recursive structure processing. In R. Lippmann, J. Moody, & D. Touretzky (Eds.), *Advances in neural information processing systems* (Vol. 3). Morgan-Kaufmann. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/1990/file/8cb22bdd0b7ba1ab13d742e22eed8da2-Paper.pdf`

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications* (Vol. 3). Springer.

McCarthy, J. (1960). Recursive functions of symbolic expressions and their computation by machine, Part I. *Commun. ACM*, *3*(4), 184–195. doi: 10.1145/367177.367199

Meier, C. (2023). *Vector symbolic architectures in Clojure.* Retrieved from `https://youtu.be/j7ygjfbBJD0` (Clojure/Conj 2023)

Neubert, P., Schubert, S., & Protzel, P. (2019). An introduction to hyperdimensional computing for robotics. *KI - Künstliche Intelligenz*, *33*(4), 319–330. doi: 10.1007/s13218-019-00623-z

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(2), 135-183. doi: 10.1016/S0364-0213(80)80015-2

nLab authors. (2024, February). *cartesian closed category.* `https://ncatlab.org/nlab/show/cartesian+closed+category`. (Revision 42)

Ororbia, A. G., & Kelly, M. A. (2023). Maze learning using a hyperdimensional predictive processing cognitive architecture. In B. Goertzel, M. Iklé, A. Potapov, & D. Ponomaryov (Eds.), *Artificial general intelligence* (pp. 321–331). Cham: Springer International Publishing. doi: 10.1007/978-3-031-19907-3_31

Plate, T. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623-641. doi: 10 .1109/72.377968

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., ... Hochreiter, S. (2020). *Hopfield networks is all you need.* arXiv. doi: 10.48550/ARXIV.2008.02217

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1), 159-216. doi: 10 .1016/0004-3702(90)90007-M

Smolensky, P., McCoy, R., Fernandez, R., Goldrick, M., & Gao, J. (2022). Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems. *AI Magazine*, *43*(3), 308-322. doi: 10.1002/aaai .12065

Stewart, T. C., & West, R. L. (2006). Deconstructing act-r. In *Proceedings of the seventh international conference on cognitive modeling* (Vol. 1, pp. 298–303).

Tamkin, A., Taufeeque, M., & Goodman, N. D. (2023). *Codebook features: Sparse and discrete interpretability for neural networks.* arXiv. doi: 10.48550/arXiv.2310.17230

Turing, A. M. (1950). Computing machinery and intelligence. , *59*, 433–460. Retrieved from `http://cogprints .org/499/`

Widdows, D., & Cohen, T. (2014, 11). Reasoning with vectors: A continuous model for fast robust inference. *Logic Journal of the IGPL*, *23*(2), 141-173. doi: 10.1093/jigpal/ jzu028

Developing and Evaluating a Computational Cognitive Model of Sensorimotor Grounded Action Selection Based on Eye-movement Behavior

Nils Wendel Heinrich[1,2], Annika Österdiekhoff[3], Stefan Kopp[3], Nele Russwinkel[1]

nils.heinrich@uni-luebeck.de

[1]Institute of Information Systems (IFIS), Department of Computer Science and Engineering, Universität zu Lübeck, Lübeck, Germany

[2]Cognitive Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany

[3]Social Cognitive Systems, Faculty of Technology, CITEC, Bielefeld University, Bielefeld, Germany

Abstract

Sensorimotor grounding of cognitive processes may be the key to why humans exhibit efficient goal-directed behavior in a variety of dynamic environments. Modeling such behavior computationally poses a challenge as the model has to exhibit equally dynamic motor control in order to ground cognitive processes in it. Once the computational model has been developed, the next challenge lies ahead: how to evaluate the model behavior using human data? Here we present an eye tracking experiment to investigate action control in dynamic environments in which fixational and smooth pursuit eye movements reflect cognitive processes of action selection. Slightly increased uncertainty in motor control leads to more cautious action selection shown by gaze being allocated closer to a reference point, whereas strongly increased uncertainty leads to the need to monitor the environment for potential threats and thus greater distances to the reference point. We equip a computational model with the hypothesized action selection processes and single out the central parameter within its structure. In the last section, a likelihood method is discussed that can be used to evaluate the model based on human eye movement behavior and to infer the parameter value.

Keywords: Sensorimotor grounding; Situated cognition; Eye-movement control; Computational modeling; Parameter inference

Introduction

Humans are incredibly efficient at pursuing similar goals in different environments and situations, such as reaching a certain position when the ground is slippery vs. when it is not. This is known as situated action control, where behavior is exerted to pursue a general action primitive (Vera & Simon, 1993). The behavior itself however can be vastly different. On slippery ground, the feet might be lifted only slightly to place them back down on the ground, gaining as much traction as possible. Intermediate goals could also be planned, such as getting to an object that we can hold on to. That may get us to the desired position safely. On the other hand, on normal ground we do not have to mind slipping. The walk to the desired position could be a long stride, almost a leap, as we can now focus our resources on getting there quickly.

Theoretical Background

Goal-directed behavior is comprised of several levels of action control. Cognitive processes are applied to conceptualise an action goal, the desired state of what the environment should look like after the next action (Kahl, Wiese, Russwinkel, & Kopp, 2022). To do this, various possible states that lead to the goal must be weighed against each other. A final selected state is then implemented by the execution of simple motor regulatory control in that motor actions continue to regulate the current state until the action goal is reached. The cognitive processes, which are responsible for action selection, are informed by the performance of motor control. It relates to how well motor actions can bring about the action goal. Based on this feedback, the action selection process can be adapted to take into account the actual motor ability to act in the current situation. By adapting cognitive processes, higher-level cognitive control is applied. The result is a hierarchical structure that exerts regulatory control at a lower motor control level and an upper cognitive control level (Badre & Nee, 2018; Kahl et al., 2022).

The behavior of agents by means of such a hierarchical structure would be based on environment or rather the ability to act within the given environment: the behavior is situated. This implies that human behavior is always improvised, at least to some degree (Clancey, 1997).

The Sense of Control in Situated Action

The ability to act successfully within an environment is assessed using an internal comparator model (Synofzik, Vosgerau, & Newen, 2008). Each motor command simultaneously generates a prediction of the effects of the command. The prediction refers to a variety of anticipated sensory inputs including proprioceptive information, perception within the own body. Here we concentrate on visual information, that refers to what is expected to be seen in the next moment. This sensorimotor prediction is matched by the comparator model with what is actually perceived visually. Both inputs will never match completely, but a simple threshold, a *sensitivity range*, can determine a magnitude of deviation that leads to a prediction error. The assumption of Kahl et al. (2022) is that these prediction errors are captured in a metric at the motor control level, the low-level Sense of Control (LL SoC). Accordingly, each prediction error results in a decrease of the LL SoC according to the size of the deviation. If the LL SoC falls below a certain threshold, this will influence an equivalent metric at the cognitive control level, the high-level Sense of Control (HL SoC), which will decrease. The HL SoC influences the action selection process and thus enables efficient, goal-directed behavior. The authors discuss that when participants are asked about their sense of control after having performed a motor task, this HL SoC will re-

flect the participants' responses. The threshold that the LL SoC has to undercut to affect the HL SoC (CCL threshold), is appropriately called the *awareness boundary*, as passing the awareness boundary implies that one becomes aware of their own insufficient motor control. By implementing two separate thresholds within the theoretical model, the authors make very specific predictions about the time trajectory of cognitive control. Only exceeding the sensitivity range several successive times or exceeding it a single time but to a large degree leads to the overcoming of the awareness boundary and thus to the activation of the cognitive processes responsible for action selection. This means that performing a motor task within a dynamic experimental environment that constantly triggers prediction errors should enable to measure when exactly cognitive control is exerted by identifying changes in behavior within the time course of a trial. But first we have to find out what effects cognitive control has on the behavior of agents. To do this, we can simply compare the behavior of agents acting within environments that evoke different amounts of prediction errors.

An Experiment Investigating the Use of Eye Movements in Situated Action Control

We conducted a study to test the various assumptions implicit to the concept of situated action control in humans. Based on the simulation environment of Kahl and colleagues, we have developed an experimental environment that we have already used for previous iterations of a cognitive computational model, the Dodge Asteroids environment (Heinrich, Russwinkel, Österdiekhoff, & Kopp, 2023). The task is to steer a spaceship through a level that is enclosed by walls on both sides. Obstacles, comets, are scattered throughout the level. The objective is to avoid crashing into either the walls or the obstacles and to cross a finish line at the bottom of the level. Similar to 2D console games, the whole level cannot be seen, but only a small observation window (Figure 1). Automatic downward movement (free fall) is induced. This means that obstacles appear at the bottom of the screen, move vertically across the screen and then disappear again at the top of the screen. Participants steer the spaceship horizontally by pressing either the Y key (step to the left) or the M key (step to the right) on the keyboard. The spaceship is fixed in the center of the screen at all times, so pressing keys causes the environment around the spaceship to move. 27 Participants played a total of 6 different levels of the Dodge Asteroids environment. For each level, the positions of the obstacles were drawn anew from a uniform distribution, which was bounded by the width and length of the level. Each level was played with 3 different intensities of input noise. Input noise has a direct effect on the efficiency of motor control influencing the ability to generate predictions for the outcome of a key press. During each game frame (the environment runs on 60FPS) in which a key for horizontal movement is pressed, the shift of the spaceship in the respective direction is drawn from a normal distribution centered over a shift of 6 pixels. We de-

fined different standard deviations of the normal distribution, in order to vary the intensity of input noise. Here we specified the 3 standard deviations 0 vs. 0.5 vs. 1, whereas the standard deviations also refer to pixels (a standard deviation of 0 meaning no uncertainty in motor control). An unexpected shift in the opposite direction of the input can also occur and does so more frequently with increased input noise.
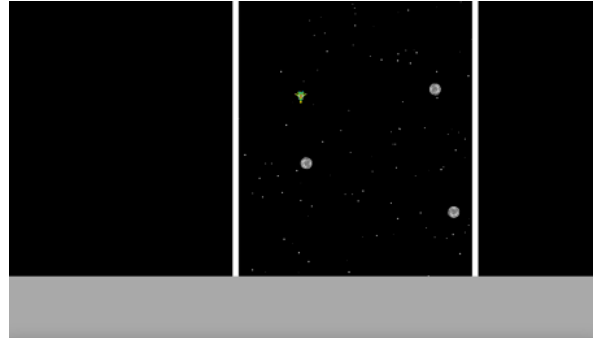
Figure 1: Depiction of an instance within the *Dodge Asteroids* environment. An observation window is shown, the space of the environment that is drawn on the screen during the experiment. Obstacles are scattered between the two walls on the left and right. The grey bar at the bottom of the screen prevents participants from looking outside the screen when they intend to look at obstacles that appear at the bottom.

We wanted to prevent participants from familiarizing themselves with certain locations (constellations of obstacles) within the levels and therefore specified a finite number of attempts. For each combination of level and input noise intensity, participants were given 3 attempts. First, all combinations were put in a list in random order and played accordingly in succession. In the event of a crash, the combination was reinserted at a random position in the list again. If all 3 attempts were used up, the combination was removed from the experimental procedure.

The screen on which the experiment was presented was placed in a distance of 80cm in front of the participants and their gaze was tracked the whole time. We used a high-frequency eye tracker, the TRACKPixx 3 with a sampling rate of 2000Hz (VPixx Technologies, Saint-Bruno, QC, Canada). A grey bar is displayed at the bottom of the screen at all times during the experiment. This prevents participants from looking outside the screen in anticipation of new obstacles and the eye tracker potentially losing the signal of the fovea.

Gaze Allocation Reflects Action Goals and the Effect of Uncertainty in Motor Control

The model of Kahl et al. (2022) assumes that although motor control is exercised continuously, it is driven by individual action goals. According to this, however, simply measuring keystrokes of participants while they steer the spaceship would not provide any information about the cognitive control of action goals. But our body has a rich sensorimo-

tor system which also includes a visual system. It has been shown how our gaze guides other manual movements (for example in natural driving; Marple-Horvat et al., 2005; Wilson, Stephenson, Chattington, & Marple-Horvat, 2007). We argue that where we move our gaze reflects the action goal we are pursuing at that point in time. Admittedly, eye movement control is also exerted continuously but this continuous stream of data can be divided into individual events: fixational and smooth pursuit eye movements. We try to equate these events with the individual action goals adopted by the cognitive model of (Kahl et al., 2022).

We applied a velocity-based detection algorithm (Engbert & Kliegl, 2003; Engbert & Mergenthaler, 2006) to filter out saccades in the eye-movement data obtained from the study described above. Subsequently, we specifically isolated fixational eye movements that showed characteristics we deemed inherit to action goals. First, they have to be initiated with the agent in peripheral vision, with a minimum distance of $5°$ (Millodot, 2014) between agent and exact gaze location. For this we computed the Euclidean distance between both points given by x and y coordinates referring to positions on the screen and converted the distance to visual degrees accounting for the distance between participants' eyes and screen. Second, the fixations had to be located in empty space, not foveating obstacles or walls. They had to remain at the relative point within the environment, following the movement of the environment on the screen, and converging on the agent over time, thus technically being categorized as smooth pursuits. This means that during these smooth pursuits, motor control was exerted specifically with the purpose of bringing the agent closer to the foveated location. We used the above conditions to create a filter that we applied to the entire data set of all fixational and smooth pursuit eye movements. What remained were all the eye movements that reflected action goals. This final data set, called *foveated action goals*, contained a total of 36,586 fixations and smooth pursuits across all 3 different intensities of input noise. Following the assumptions we already integrated into the earlier iteration of our computational cognitive model (Heinrich et al., 2023), we investigated whether participants initiated action goals closer to the agent the less control they had over it. This translated into the hypothesis that the distance of foveated action goals to the agent decreases with increasing intensity of input noise.

We applied linear mixed modelling using *Julia* 1.9.3 (Bezanson, Edelman, Karpinski, & Shah, 2017) and the *MixedModels* package for statistical modeling (Bates, 2015) to test our hypothesis[1]. We used a random seed, the MersenneTwister(36) of the Random module. A box-cox distributional analysis (Box & Cox, 1964) of the fixational distance to the agent indicated a transformation to the logarithmic scale. Note that therefore all β-values and CI bounds are reported on the logarithmic scale. We further explored the random effects structure of the model by referring to the Bayesian infor-

---

mation criterion for model selection (Chakrabarti & Ghosh, 2011). The final model selected model included random intercept effects for participant ID and the number of visible drift tiles. The latter being a manipulation of the experiment that not of interest when it comes to analyzing the hypothesis described here. We found that the distance to the agent in visual degrees decreased with weak input noise compared to no input noise ($β = −0.026$, $σ = 0.005$, $CI_{95\%} = [−0.035, −0.017]$, $p < .001$), supporting our hypothesis. However, contrary to our hypothesis the distance increased with strong input noise compared to weak input noise ($β = 0.039$, $σ = 0.005$, $CI_{95\%} = [0.030, 0.049]$, $p < .001$). Foveated action goals were thus located closer to the bottom of the screen.

Why is it that with a higher loss in motor control, action goals are suddenly planned further ahead? We derived new hypotheses that consider the role of bottom-up processing in visual perception. Fixational eye movements fulfil several functions. They are not only to maintain the top-down action intention and perceive the effects of motor control while pursuing the action goal. They certainly are also used for active perception of the environment and to monitor specific locations. Therefore, the final gaze location is a compromise of the location of the action goal and the need to sample the bottom edge of the screen for incoming obstacles. Under enormous motor control loss, the need to detect incoming obstacles as early as possible to ease cognitive load when planning paths through said obstacles becomes more important. It may even outweigh the need to foveate the action goal.

In the following, we will set up a computational cognitive model in which we precisely define our hypotheses. In this paper, we present how to verify a model that produces dynamic eye movements using mean statistics. The advantage of a computational model would be that once it is verified, we can generate highly accurate predictions about the temporal trajectory of behavior using simulations.

A Computational Cognitive Model of Situated Action Control

We have already successfully implemented several hypotheses in a previous iteration of a computational cognitive model (Heinrich et al., 2023) in the following referred to as two-layer architecture to emphasise its body structure. We have also improved and expanded on the model's internal processes. The current version of the two-layer architecture used in this paper is implemented in Python (Van Rossum, Guido & Drake, 2009) and its internal functions are described below.

The two-layer architecture engages in situated action control by means of a rather straightforward action selection process. In order to effectively identify action possibilities within the immediate environment, a convolution of the visual input is generated and held at the cognitive control level. The visual input is the part of the Dodge Asteroid environment that is inside the observation window between the walls and below the agent. All pixels of the visual input are assigned to

one of 56 pools of equal size. The number of pools (N_{pools}) is a free parameter that reflects the granularity of the representation of the visual environment and can be chosen from a defined set of values (20, 30, 42, 56, 72; a higher value refers to a higher granularity). The pools are arranged in a convolution grid with 6 rows and 7 columns. Within a given pool the color activation of each pixel, comprised of its RGB values bounded between 0 and 1, is factored into an overall mean activation of the pool. If the mean activation exceeds a threshold of 0.03 (meaning that if 3% of the pixels within the pool are colored), the pool is considered populated by an obstacle and eliminated from the set of action possibilities (Figure 2). Subsequently all other pools of the *same* column that are *below* the rejected pool are eliminated, reflecting that this horizontal location is not safe due to the vertical movement of obstacles (the subsequently eliminated pools might have even been unpopulated). All other unpopulated pools are accordingly identified as possible locations for an action goal. Regarding selecting the horizontal location of the action goal, a column is chosen based on the minimum distance to the current position of the agent. This is grounded in our model applying a heuristic that is meant to minimise the need for motor control. The row of the final action goal (vertical location) is determined using Equation 1:

$$\text{row}_{\text{action goal}} = \lfloor \text{HL SoC} * N_{rows} \rfloor, \quad (1)$$

with N_{rows} being the total number of rows in the convolution grid (in our example N_{rows}=7, but it changes with N_{pools}). It factors the HL SoC that ranges from 0 to 1. Of the final selected pool, the exact center is specified as intended location and thus the top-down action goal. It might happen that there are no available pools in the corresponding row under the current granularity. In this case, the action field is generated again, but this time under the highest granularity (N_{pools}=72). The larger number of pools and thus the finer differentiation of the visual space makes it possible to identify free spaces between obstacles where previously there were none. We consider the resampling of the action field as hacking because it has no theoretical justification but allows the model to always act on the basis of an action goal. However, one could argue that the more obstacles there are in the visual environment, the more accurate the visual environment is searched for free spaces to steer to. The action goal is passed to the motor control layer for implementation. At this level, the final position of the action goal is biased by bottom-up processing of color activation of individual pixels that may attract the gaze. A two-dimensional probability density distribution with all density gathered within the selected pool is integrated with a heatmap of color activation over all pixels in the visual input in Bayesian fashion. Of the resulting two-dimensional probability density distribution, the point of highest probability density is selected as the final action goal and the model will initiate a fixation at that exact location (Figure 2), the foveated action goal. It is utilized for online motor control when navigating the environment.
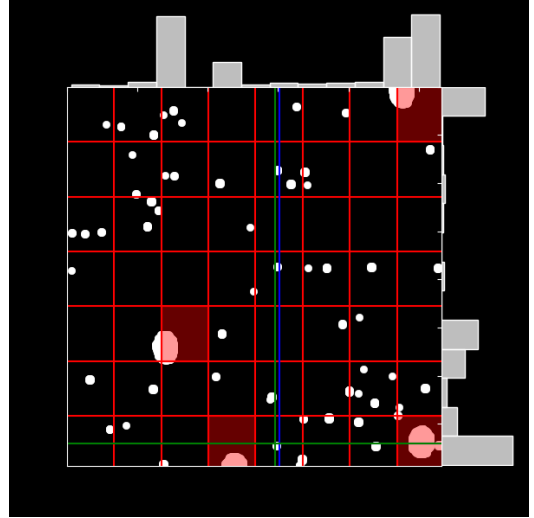
Figure 2: Illustration of the grid with N_{pools}=56 and color activation within the visual input of the two-layer architecture. The bars on the axes indicate the locations of high activation. The center of the selected pool (top-down action intention) is indicated by where the blue lines meet. The final action goal (compromise between top-down action intention and bottom-up color activation) is shown at the point where the green lines meet.

In each frame of the game environment, the two-layer architecture will update the location of the foveated action goal by the inferred movement of the visual environment. The inferred movement is the posterior of two ongoing Bayesian integration processes individually for vertical and horizontal movement of the environment. The likelihood is given by the observed movement. We introduce uncertainty to the observation by means of a noise factor, resulting in a gaussian probability density function. This represents the internal noise of the visual system. The prior is the posterior of the inference step before. The model always starts every trial with a flat prior (uniform distribution bounded by the dimensions of the observation window). The updated horizontal position of the foveated action goal is matched with the horizontal position of the agent and the direction of necessary movement is determined (left vs. right depending on the horizontal position of the agent in relation to the foveated action goal). Finally, the horizontal distance is minimized over time by maintaining the directional input as many frames as needed. This is how the model exercises motor control in the form of key presses.

If an obstacle appears whose horizontal location is inferred to be on the horizontal section of the current action goal, the action goal is abandoned, and a new action selection process is triggered.

According to the assumptions of Synofzik et al. (2008), the two-layer architecture applies predictive coding to identify prediction errors. Similar to how the model infers the movement of the environment, it infers the horizontal shift

resulting from its own motor input. It does so by repeatedly integrating the perceived shift in a Bayesian manner with the likelihood being the actual observed shift with the noise factor added every time it steers the spaceship. This way, the two-layer architecture will reach a precise estimate of its own movement within the environment resulting from exercising motor control as the posterior of the integration process at time t will be the prior for the integration process at t_{+1}. The estimate is then used to generate a feedforward prediction whenever the model exercises motor control since it anticipates the outcome of its own action. The prediction is compared to the observed shift. For the comparison, the model refers to the Kullback-Leibler divergence. If the divergence exceeds a value of 0.0001 (which represents the *sensitivity range*), a prediction error is identified and the control metric specifically for motor control, the LL SoC, is reduced by the value of the divergence. If the LL SoC falls below the threshold value of 0.3, which is the *awareness boundary* parameter of the model, this has a direct effect on the HL SoC. It is now reduced by a fixed proportion (0.25) of its own value. This reflects the assumption that under increased cognitive control, a loss of motor control also leads to an increased loss of cognitive control. Consequently, the next action goal will be chosen closer to the agent (Equation 1).

The two-layer architecture described here attempts to explain the effects of prediction-based motor control on action selection processes while accounting for visual perception and oculomotor control. In a next step, we can conduct a comparison between the mean statistics of foveated action goals of our model with those of human participants to tune parameter values within the model. The distinctive feature of the two-layer architecture is that it makes explicit assumptions about the time dependency of cognitive control. Error signals in the sensorimotor system must effectively exceed two individual thresholds before they take effect on the action selection processes. Given an adequate fit to human behavior, we are able to predict moments of control loss in humans and their implications on action selection. However, since foveated action goals are measured using highly dynamic eye movements, estimating a match between model and human data poses some challenges.

The Feasibility of Using Eye Movements to Infer Parameters Values

For a preliminary comparison, we simulated model behavior for each of the combinations of level and input noise intensity exactly once under each of the different possible values for the convolution granularity parameter N_{pools}. The two-layer architecture therefore played the same levels as the human participants. Our hypothesis focused on the distance of the gaze to the agent. Therefore, we calculated the distance for the simulated data and converted it to visual degrees, as was the case for the human data. It will be the only summary statistic we assess in this work but note that many more statistics such as the duration of fixations and smooth pursuits or

the properties of saccades might be assessed simultaneously.

Applying kernel density estimation (kernel density estimate; KDE), we can visually compare the probability distribution of the distance from the foveated action goal to the spaceship between model and human data. High peaks indicate that foveated action goals were initiated at this distance from the agent particularly often. Figure 3 shows the individual KDEs of the model data generated with the N_{pools} parameter value equal to three different values of the defined set split each for the three different input noise intensities. We can see immediately that for each input noise intensity the point of highest probability density for the model is always shifted towards lower distances compared to human data. The model data under N_{pools}=56 yielded the best visual fit (Figure 3 b). Lower parameter values also showed a tendency towards smaller distances (for reference N_{pools}=20; Figure 3 a). This might be due to the fact that with increased HL SoC, the last vertical row of the grid is selected. Compared to coarser grids, with a finer grid the center of the individual pools in the last row is further away from the agent resulting in greater distances to the agent. Further increasing the granularity of the convolution grid, N_{pools}=72 resulted in foveated action goals with less variance (Figure 3 c). The distinct peak under this parameter value has a worse visual fit to the human data compared to N_{pools}=56.
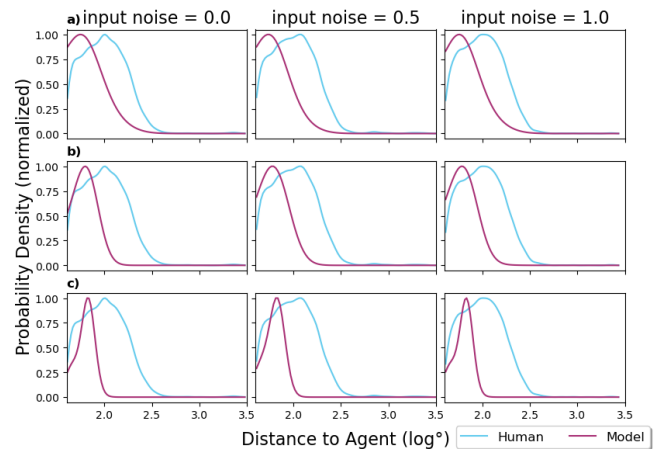
Figure 3: Normalized kernel density estimates for our exemplary metric for human data and model data simulated under three different values for N_{pools}. The specific values were: a) N_{pools}=20, b) N_{pools}=56, and c) N_{pools}=72. The metric is individually compared within the three different input noise intensities, with column 1 showing KDEs within input noise 0.0, column 2 within input noise 0.5, and column 3 within input noise 1.0.

The Log Synthetic Likelihood

The granularity of the convolution grid turns out to be a powerful parameter of the model, probably even the most important one to fit to human data. But conducting a parameter

inference on the basis eye-movement data poses a challenge. We need an objective method that reflects the fit across all three different input noise intensities simultaneously and outputs a single value that quantifies the overall agreement between model and human data. Wood (2010) proposed such a method in the construction of a synthetic likelihood. It is a phase-insensitive method that attempts to discard noise in the data and to consider only the important dynamics. This might prove helpful in the assessment of noisy eye movement data. We will repeatedly change the value for N_{pools} and compute the synthetic likelihood as depicted by Wood (2010) for the data simulated under the specific parameter value. Based on the assumption that humans elicit behavior according to our model, the summary statistics s of the human data y can be recovered simulating under unknown model parameter values θ (here we will only vary N_{pools}). As mentioned above, the single summary statistic used here will be the distance to the agent in log visual degrees for the three individual input noise intensities. We will therefore assess the fit for three values simultaneously. Using the model, a number of Nr data sets $(y_1^\star, y_2^\star, y_3^\star, ..., y_{N_r}^\star)$ are generated for each different value of θ (here we set Nr equal to 1). These data sets are reduced to summary statistics $(s_1^\star, s_2^\star, s_3^\star, ..., s_{N_r}^\star)$ as was the human data y (with each $s^\star$ comprising of the three different distance to the agent statistic for the various input noise intensities). Next, we compute the mean vector $\hat{\mu}_\theta$ (Equation 2) and the covariance matrix $\hat{\Sigma}_\theta$ (Equation 3).

$$\hat{\mu}_\theta = \sum_i \frac{s_i^\star}{N_r} \qquad (2)$$

$$\hat{\Sigma}_\theta = \frac{SS^T}{N_r - 1} \qquad (3)$$

Where $S = (s_1^\star - \hat{\mu}_\theta, s_2^\star - \hat{\mu}_\theta, ...)$. $\hat{\mu}_\theta$ and $\hat{\Sigma}_\theta$ are given as input into Equation 4 from which we finally obtain the synthetic likelihood.

$$l_s(\theta) = -\frac{1}{2}(s - \hat{\mu}_\theta)^T \hat{\Sigma}_\theta^{-1}(s - \hat{\mu}_\theta) - \frac{1}{2}log|\hat{\Sigma}_\theta| \qquad (4)$$

This procedure is repeated for every different N_{pools} value θ. After obtaining all the synthetic likelihood values, we can plot a likelihood profile with its peak indicating the θ under which the human data y was most likely generated (Figure 4). The likelihood profile peaks at N_{pools}=56 which agrees with our visual assessment[2].

Limitations

In this first application of Wood's (2010) log synthetic likelihood method, we successfully identified a parameter value for the granularity of the convolution grid that might underlie human action selection. This shows that parameter inference is possible even on the basis of extremely noisy eye-movement data.

---

[2]Original code and data used to generate plots and obtain likelihood values can be accessed via the link: https://github.com/nilsheinrich/ICCM_2024.git
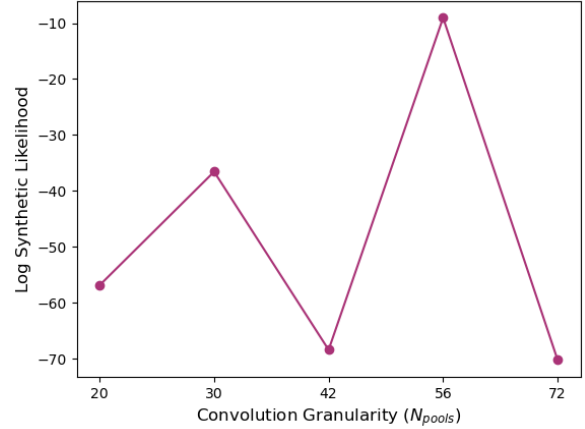
Figure 4: Likelihood profile showing the individual results of the log synthetic likelihood method for determining the fit of model and human data (y-axis), for every value of the convolution granularity parameter N_{pools} which was used to simulate the model data (x-axis).

The log synthetic likelihood method would indicate a best possible fit of model and human data if it approaches the value 0. This does not apply to the likelihood values obtained here and therefore we can still consider the model to be inadequate. We aim to strengthen our confidence in the likelihood method by increasing the number of simulations N_r to obtain a more robust estimate for the mean vector $\hat{\mu}_\theta$ and covariance matrix $\hat{\Sigma}_\theta$. At the same time, as already mentioned, further summary statistics could be assessed.

The following specific adjustments could be made to the two-layer architecture. The vision of the model is not foveated yet. The general noise factor in the observation should be extended to grow with increasing distance to the foveated location. This way the locations of objects far away from the foveated action goal would be even more uncertain and this might generate the need to shift gaze to these specific objects to better estimate their locations. This is probably what is seen in the human data. Participants execute fixations specifically to pinpoint objects that just appeared. This might happen especially often under decreased HL SoC. However, we miss this capability in our model, as it uses visual attention exclusively to foveate action goals.

Lastly the integration process to estimate environmental movement or positional shift due to own actions might be tuned to better reflect the general assumptions in literature. Here when updating the estimate new observations should be weighted lower if no prediction error was detected, reflecting that these reafferent signals that confirm predictions hold little new information (Fiehler, Brenner, & Spering, 2019; Shoshina et al., 2020).

Acknowledgements

References

Badre, D., & Nee, D. E. (2018, February). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences*, *22*(2), 170–188. Retrieved 2024-02-26, from https://www.sciencedirect.com/science/article/pii/S1364661317302450 doi: 10.1016/j.tics.2017.11.005

Bates, D. (2015). *MixedModels: A Julia Package for Fitting Statistical Mixed-Effects Model.* Retrieved from https://github.com/dmbates/MixedModels.jl

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017, January). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. Retrieved 2024-01-03, from https://epubs.siam.org/doi/10.1137/141000671 doi: 10.1137/141000671

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *26*(2), 211–243. Retrieved 2024-01-03, from https://academic.oup.com/jrsssb/article-abstract/26/2/211/7028064 (Publisher: Oxford University Press)

Chakrabarti, A., & Ghosh, J. K. (2011, January). AIC, BIC and Recent Advances in Model Selection. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of Statistics* (Vol. 7, pp. 583–605). Amsterdam: North-Holland. Retrieved 2024-01-03, from https://www.sciencedirect.com/science/article/pii/B9780444518620500186 doi: 10.1016/B978-0-444-51862-0.50018-6

Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge university press.

Engbert, R., & Kliegl, R. (2003, April). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. Retrieved 2024-04-22, from https://www.sciencedirect.com/science/article/pii/S0042698903000841 doi: 10.1016/S0042-6989(03)00084-1

Engbert, R., & Mergenthaler, K. (2006, May). Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences*, *103*(18), 7192–7197. Retrieved 2024-04-22, from https://www.pnas.org/doi/full/10.1073/pnas.0509557103 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.0509557103

Fiehler, K., Brenner, E., & Spering, M. (2019). Prediction in goal-directed action. *Journal of Vision*, *19*(9), 10–10. Retrieved 2024-05-28, from https://www.arvojournals.org/article.aspx?articleid=2748974 (Publisher: The Association for Research in Vision and Ophthalmology)

Heinrich, N. W., Russwinkel, N., Österdiekhoff, A., & Kopp, S. (2023, July). A Straightforward Implementation of Sensorimotor Abstraction in a Two-Layer Architecture for Dynamic Decision-Making. In *MathPsych/ICCM/EMPG 2023*.

Kahl, S., Wiese, S., Russwinkel, N., & Kopp, S. (2022, March). Towards autonomous artificial agents with an active self: Modeling sense of control in situated action. *Cognitive Systems Research*, *72*, 50–62. Retrieved 2023-11-27, from https://www.sciencedirect.com/science/article/pii/S1389041721000851 doi: 10.1016/j.cogsys.2021.11.005

Marple-Horvat, D. E., Chattington, M., Anglesea, M., Ashford, D. G., Wilson, M., & Keil, D. (2005, June). Prevention of coordinated eye movements and steering impairs driving performance. *Experimental Brain Research*, *163*(4), 411–420. Retrieved 2024-01-25, from https://doi.org/10.1007/s00221-004-2192-7 doi: 10.1007/s00221-004-2192-7

Millodot, M. (2014). *Dictionary of optometry and visual science e-book.* UK: Elsevir Health Sciences. (p. 250)

Shoshina, I., Isajeva, E., Mukhitova, Y., Tregubenko, I., Khan'ko, A., Limankin, O., & Simon, Y. (2020). The internal noise of the visual system and cognitive functions in schizophrenia. *Procedia Computer Science*, *169*, 813–820. Retrieved 2024-05-28, from https://www.sciencedirect.com/science/article/pii/S1877050920302817 (Publisher: Elsevier)

Synofzik, M., Vosgerau, G., & Newen, A. (2008, March). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–239. Retrieved 2023-11-27, from https://www.sciencedirect.com/science/article/pii/S1053810007000268 doi: 10.1016/j.concog.2007.03.010

Van Rossum, Guido, & Drake, F. L. (2009). *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.

Vera, A. H., & Simon, H. A. (1993). Situated Action: A Symbolic Interpretation. *Cognitive Science*, *17*(1), 7–48. Retrieved 2024-02-26, from https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1701_2 doi: 10.1207/s15516709cog1701_2

Wilson, M., Stephenson, S., Chattington, M., & Marple-Horvat, D. E. (2007, January). Eye movements coordinated with steering benefit performance even when vision is denied. *Experimental Brain Research*, *176*(3), 397–412. Retrieved 2024-01-25, from https://doi.org/10.1007/s00221-006-0623-3 doi: 10.1007/s00221-006-0623-3

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, *466*(7310), 1102–1104. (Publisher: Nature Publishing Group)

Exploring Memory Mechanisms Underlying the Continued Influence Effect

[1]**Alexander R. Hough (alexander.hough.1@us.af.mil)**
[2]**Othalia Larue (othalia.larue@parallaxresearch.org)**

[1]Air Force Research Laboratory, Wright-Patterson AFB
[2]Parallax Advanced Research, Beavercreek, OH 45431 USA

Abstract

Humans have cognitive vulnerabilities that can be leveraged to influence individuals. One such vulnerability is the continued influence effect (CIE), where misleading information can have a lasting effect even after corrections or factual discrediting information is presented. The CIE has been addressed experimentally and memory-based explanations exist. Currently, no cognitive models exist to specify mechanisms for prediction, simulation, and detailed testing of hypotheses. Here, we discuss relevant literature and propose a novel cognitive model implemented in ACT-R to investigate memory mechanisms underlying the CIE. We demonstrate the utility of our initial model using simulations which show how the CIE emerges from memory processes and discuss plans for future research.

Keywords: ACT-R; Cognitive modeling; misinformation; continued influence effect; knowledge representation

Introduction

Humans have systematic cognitive biases (e.g., confirmation bias and loss aversion) that could be exploited to influence attitudes or judgment (Kahneman, 2011). Misinformation leverages those biases, and due to technology and social media advancement, is now more prevalent (Wang et al., 2019). An example of bias increasing susceptibility to misinformation is the continued influence effect (CIE) (Johnson & Seifert, 1994; Lewandowsky et al., 2012), where misinformation has a lasting effect on decisions despite the presentation of corrections or discrediting facts. Experimental manipulations demonstrated the CIE is robust and, at best, can be reduced by 50% (Lewandowsky et al., 2012). Recently, experts in the field (Ecker et al., 2022) stated the literature lacks: 1) a psychology of misinformation, 2) realistic scenarios in experiments, 3) an understanding of the interplay between cognition, social, and emotional factors, and 4) an overarching theory and model including these factors and spanning from individuals to groups. Here, we present our initial effort as we start addressing gap 3 with a cognitive model to explain memory mechanisms underlying the CIE.

Explanations of the CIE focus on episodic memory, where a piece of information entering long-term memory (LTM) cannot be erased but rather re-activated or associated with different information (Wilkes & Leatherbarrow, 1988). This can lead to memory errors related to competing memory activations (Ayers & Reder, 1998; Ecker et al., 2010), recency effects (Ecker et al., 2015), or familiarity-based fluency (Ecker et al., 2011; Swire et al., 2017). In addition, more available or easier-to-access information tends to be weighted more in judgment (Kahneman, 2011). Memory activation and availability present an interesting issue with corrections. They may re-activate misinformation by repeating some elements, leaving it more "available" than the correction. Corrections may not fit into the coherent mental model constructed during the presentation of the original (mis)information (Wilkes & Leatherbarrow, 1988; Johnson & Seifert, 1994; Lewandowsky et al., 2012). This may result in fewer retrieval pathways to the correction (Seifert, 2002; Mayo et al., 2004). Gilbert et al. (1990) suggest retractions require effective negation "tags" to overcome these memory errors and successfully retrieve the corrected information (Johnson & Seifert, 1998). However, corrections or negation may not be effective if they do not fit well enough into the relational or causal structure between memories to replace the misinformation (Johnson & Seifert, 1994).

The CIE is largely a memory phenomenon that interacts with other cognitive processes. The literature provides hypotheses focused on specific rather than holistic aspects of memory and lacks cognitive models to thoroughly test and compare hypotheses. To better understand how cognitive processes interact, make predictions, and explain behavior, we leverage the ACT-R cognitive architecture (Anderson, 2007). We previously simulated the CIE (Hough et al., 2023) by providing misinformation to a model that learned to make decisions based on sets of cues in a binary decision making task (Halverson et al., 2018; Myers et al., 2015). Here, we explored the memory mechanisms underlying the CIE with a more appropriate task (i.e., CIE experiment). During this process, we faced several challenges with parsing language, methods used to analyze data from the literature, and approximating behaviors like question answering and beliefs. We present our contributions toward addressing these challenges, our novel CIE model, and areas for improvement.

The CIE Task

CIE experiments typically use two stories about scenarios, where the first contains misinformation and the second has the correction. The general finding is that corrections reduce, but do not eliminate inferences consistent with misinformation (Brydges et al., 2018; Ecker & Antonio, 2021; Ecker et al., 2017; Johnson & Seifert, 1994). We focused on six scenarios from a previous experiment (Ecker & Antonio,

2021) available on OSF: `https://osf.io/awsf5` (see Table 1). Rather than using separate narratives for each scenario, the study used one narrative presented one sentence at a time with the critical (mis)information as the second sentence and retraction (i.e., correction) as the second-last sentence. Retractions included a source and their profession. Source information was used to manipulate credibility (low and high) and trustworthiness (low and high). There were also two control conditions: best possible retraction and no retraction.

Table 1: Critical (mis)information and retractions for the six scenarios from Ecker and Antonio (2021)

Scenario	Critical (mis)information	Retraction
Football scandal	Larsson is believed to have tested positive for performance enhancing drugs	Oliver Lindgren stated that "I do not believe that Larsson has engaged in drug use
Recreational fishing	There have been reports that since the introduction of the restrictions, marine tourism numbers have grown substantially as more recreational snorkelers and divers have been attracted to the state	Barry James has stated that 'marine tourism numbers have certainly not increased since the introduction of the restrictions
Contaminated water	It is believed that the fish deaths were caused by contamination with industrial pollutants from a nearby mining company	Todd Hunter explained that "there was no contamination from mining operations."
Food additive	The list contains many food additives that have been suggested to pose serious health risks, including increased risk of cancer and ADHD	Randall Carter stated that "all food additives on the list have been thoroughly tested and are safe for human consumption."
Inflammatory joint disease	Symptoms of inflammatory joint conditions can be treated effectively through remedial yoga.	Debra Phillips has stated that effective treatments will almost always include pharmaceutical intervention, as practices such as yoga are not effective.
Anti-viral drug	A new anti-viral drug that was promising a breakthrough in treatment of viral infections is being withheld because of safety concerns.	Gerard Hintzman stated that "despite the delay of the market launch, there are no safety concerns regarding Nanofadol.".

After a 10 minute distractor task, participants completed two questionnaires including: 1) a recall question and inferential reasoning questions relating to each scenario, and 2) belief ratings for the critical (mis)information, and retraction on a 10-point scale (not at all-very strong). The first experiment showed a slight reduction in the CIE (i.e., reliance on misinformation) in conditions with higher trustworthiness. In addition, the critical (mis)information was rated as more believable than retractions. The authors also mentioned that retractions provided no context or supportive arguments, which rules out the mental model explanation. Here, we attempt to capture the results for both questionnaires.

Cognitive Model of the CIE

We implemented our CIE model within the ACT-R cognitive architecture (Anderson, 2007). ACT-R is a hybrid cognitive architecture with symbolic and sub-symbolic structures, and modules representing systems of the mind. The CIE model uses the goal, vision, imaginal, and both declarative and procedural memory modules. The goal module serves as the model focus and stores goal-relevant information. The vision module allows the model to perceive visual stimuli, and the imaginal module serves as temporary working memory. Declarative memory stores information as chunks and captures memory dynamics. The procedural module uses condition-action rules (i.e., productions) to represent knowledge about how to do things and to drive the model's behavior. Given the nature of the CIE, here we focus on declarative memory.

ACT-R Declarative Memory
CIE research suggests memory for misinformation and corrections compete. In ACT-R there are existing mechanisms in declarative memory capable of capturing this competition through chunk activation. Chunks are the basic units of declarative memory and are comprised of slots that contain values (e.g., situation, decision, and utility). A chunk has an activation value corresponding to the probability and speed it will be retrieved in a given situation. Activation, A_i, is determined by starting or base level activation, B_i, how much activation spreads among other chunks in memory, S_i, partial matching for degree chunk matches retrieval requests, P_i, and added activation noise, ε_i. Here, we only use B_i and ε_i

$$A_i = \mathbf{B_i} + S_i + P_i + \varepsilon \quad (1)$$

The base level term, B_i, describes opposing dynamics of learning with experience and forgetting across time. It is stated as:

$$B_i = \log\left(\sum_{j=1}^{n_i} t_{ij}^{-d}\right) \quad (2)$$

where n_i is the number of times chunk i has been used or retrieved, t_{ij} is elapsed time in seconds since the jth retrieval, and $d \in [0,1]$ is a decay parameter. Therefore, if a chunk is created with a high base level activation or is used frequently, it will have more influence in decision making.

In our model, we used six declarative memory parameters. Retrieval threshold which restricts chunks lower than the threshold from being retrieved and starting base-level activation, B_i, were arbitrarily set at 1 and 10, respectively. We used the recommended value of .5 for base-level decay, d, and set activation noise, ε_i, at .25, which is on the low end of the recommended range of .2 - .8. As we are exploring appropriate methods to navigate memory representations, we took some liberty with declarative finsts by setting the number of items (i.e., amount of items in memory marked as attended) higher than the number of items in scenarios and span (i.e., time span items remain marked) high enough that finsts won't clear during knowledge representation navigation. Note that we did not include partial matching, P_i, or spreading activation, S_i, terms from the activation equation (equation 1). In our CIE model, we focus on retrievals and changes in chunk activation based on declarative memory dynamics (e.g., decay and frequency of use) to capture the competition between (mis)information and retractions.

CIE Model Knowledge

Models that rely on language processing require significant existing knowledge about words and how to extract meaning based on sentence construction. Ideally, the model would read sentences, comprehend them, and form chunks, rather than hard coding chunks based on scenarios. However, our main focus is on memory competition underlying the CIE. It is good practice to generate knowledge representations independent of the current model or modeler to reduce "tailorability" (Forbus et al., 2017) where model results depend on representation choices of the modeler. Therefore, we used Romero et al. (2023)'s method by providing an in-context learning example to ChatGPT to parse sentences and convert them into lists of predicates. We initially used the predicate structure of one scenario to build the model. After developing the model, we ran the remaining five scenarios through ChatGPT and found the output changed dramatically. Therefore, we had to manually create predicates for five scenarios. We followed the natural sentence structure and used the predicates provided from ChatGPT as a guide to avoid introducing our own bias.

Table 2: Word pairs used as chunks for food additive scenario

Type	Predicates
Neutral	(list food-additives)
Critical (mis)info	(food-additives health-risks) (health-risks serious) (serious cancer) (serious ADHD)
Retraction	(Randall-Carter source) (source statement) (statement all-food-additives-on-the-list-have-been-thoroughly-tested-and-are-safe-for-human-consumption) (food-additives tested) (tested thorough) (tested safe-for-human-consumption)

CIE Model Processes

We developed the model with both simplicity and generality in mind so that we could explore the space and address some challenges with modeling the CIE. In Ecker and Antonio (2021), participants read scenarios with misinformation and a correction (Table 1), then were asked a series of questions. Here, we explored methods for the model to answer the general open-ended recall questions (summarize the report in a sentence or two and what was the main point of the report?) and approximate belief ratings for critical (mis)information and retractions. To simulate the experiment, the model was presented with one scenario at a time in random order. Scenarios were presented as predicate pairs, given one at a time. The model is presented with all scenarios, then navigates mental representations of scenarios while preparing to answer questions (see Figure 1). The model directs visual attention to find, attend, and read (i.e., encode) predicate pairs (i.e., words). After reading one word, the model attempts to recall if this information was previously encountered and associated with another word (retrieve-assoc), and if so, it retrieves that chunk and increases its activation. After this retrieval the model finds and reads the second word. Alternatively, if the model reads one word and cannot retrieve an association,
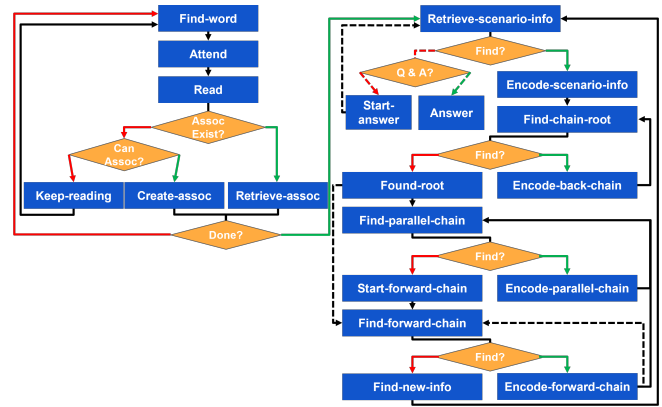
Figure 1: CIE model processes (blue rectangles), conditions (yellow diamonds), and flow of behavior (arrows).

it reads the second word (keep-reading). In either case, after reading the second word, the model creates a chunk with those two predicates (create-assoc). This process continues until all information for each scenario has been presented and stored as chunks. The model is then directed to answer questions about each scenario in random order. We implemented processes to simulate the formation of and navigation within a mental representation by chaining or linking chunks. The model starts by randomly recalling a chunk not yet retrieved (retrieve-scenario-info). If a chunk is recalled, it is encoded (encode-scenario-info) by placing it into the imaginal buffer, and the model attempts to find the root or starting link in the chain it belongs to through back-chaining (find-chain-root). Back-chaining uses the first word of the current chunk in the imaginal buffer as a cue (e.g., **food-additives** health-risks) to retrieve a chunk in memory with a matching second word (e.g., list **food-additives**). If back-chaining is successful, the recalled chunk is encoded (encode-back-chain), and the model continues back-chaining until the root is found (found-root). Once the root is found, the model checks if any parallel chains exist (find-parallel-chain). It uses the current chunk's first word as a cue (e.g., **serious** cancer) to retrive a chunk in memory with a matching first word (e.g., **serious** ADHD), and if found, it is encoded (encode-parallel-chain). If no parallel chains are found, the model initiates forward-chaining (start-forward-chain). It uses the current chunk's second word as a cue (e.g., food-additives **heath-risks**) and attempts to recall (find-forward-chain) a chunk in memory with a matching first word (e.g., **health-risks** serious). If a chunk is retrieved, it is encoded (encode-forward-chain), and the model goes back to check if a parallel chain exists. Therefore, when forward chaining, the model checks if a parallel chain exists after each retrieval. If the model fails to find a chunk while forward-chaining, the chain cycle breaks, and the model prepares to start a new chain (find-new-info) by recalling a chunk not yet attended (retrieve-scenario-info). Finding a chunk, back-chaining to the root, and forward-chaining with parallel-chain checks repeats until the model

cannot retrieve a new (i.e., not recently attended) chunk to chain. At this point, all the information for the current scenario has been recalled, and the model is ready to generate an answer (start-answer). How an answer "should" be generated is an open question. We opted to have the model retrieve the most active chunk and then answer with the most active chain it belongs to. During this process (dotted lines in Figure 1), the model retrieves the most active chunk (retrieve-scenario-info), finds the root through back-chaining (find-chain-root), skips parallel-chaining (find-parallel-chain), and forward-chains (find-forward-chain). Once it fails to forward-chain, it gives the completed chunk chain as the answer.

Results

To assess the model's performance, we simulated the same number of participants ($N = 53$) to compare with human data from Ecker and Antonio (2021). Our initial model lacks mechanisms to interpret the source information used in the experiment. Therefore, the human data was reorganized into scenarios by collapsing all source manipulations. In addition, the model cannot understand questions and give human-like responses, so we were limited in what we could compare. We targeted the open-ended questions and belief ratings in critical (mis)information and retractions. In the human experiment, five open-ended questions were each scored 0-1 depending on if critical (mis)information was not mentioned or disavowed (0), partially endorsed (.5), or fully endorsed (1), resulting in a critical (mis)information score range of 0-5. Since our model only gave a single simulated summary, we scaled the human scores by dividing by five. To approximate a critical (mis)information score for the model, we produced a score from 0-1 based on the proportion of words the model gave as the summary answer that matched words in the scenario's critical (mis)information. This score was transformed to match the 0, 0.5, and 1 values assigned by the human scorer in the original experiment by rounding up scores > 0.2 to 0.5 and scores > 0.5 to 1.

We explored how well the model captured the trend in critical (mis)information scores across scenarios using a correlation and assessed the average difference between the model and human data using root mean squared error (*RMSE*) (see Figure 2). We found the model had a non-significant negative relationship with the trends in the human data, $r(10) = -0.46, p = 0.36$, and a rather high average difference, $RMSE = 0.21$. This comparison is limited as it required modification of the human data and some liberty in approximating a similar score for the model with one answer summary. It was challenging to implement processes to approximate open-ended question answering and to calculate the critical (mis)information score with our model's summary answer. This presents an issue for future modeling work using this methodology, which is common in the CIE literature.

The comparison with belief scores was more straightforward and appropriate. In Ecker and Antonio (2021), belief ratings for critical (mis)information and retractions were
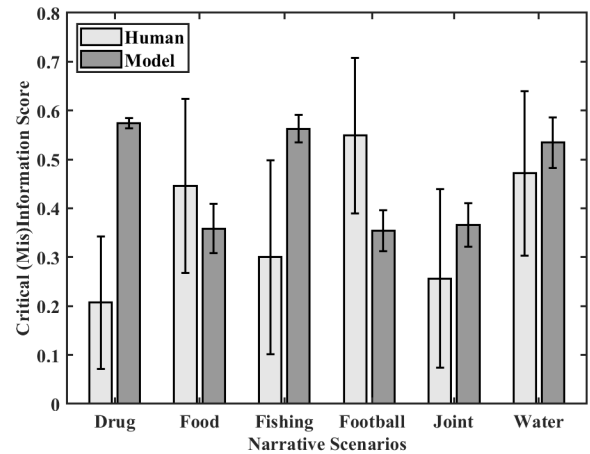
Figure 2: Critical (mis)information scores for human and model data across all six scenarios. Error bars are SEM

given on a scale from 1-10. We were interested in differences in beliefs, not differences across conditions, so we collapsed scenarios. We performed a t-test and found the same effect reported in the paper across source manipulations: belief ratings for critical (mis)information ($M = 6.86, SD = 2.05$) were higher than retractions ($M = 4.86, SD = 2.54$), $t(317) = 8.43, p = 2.34e - 15$. As a preliminary method to approximate the model's "beliefs", we used chunk activations after the model had navigated through its mental model of each scenario prior to answering the summary question. We averaged the activation across chunks representing the critical (mis)information and retractions for each scenario. The ratio of critical (mis)information and retractions chunks varied per scenario: 1) food additive was 4:6, 2) fishing was 5:5, 3) football was 3:5, 4) joint disease was 4:7, 5) contaminated water was 5:5, and 6) ant-viral drug was 7:7. Similar to the human data, we found chunk activation was higher for critical (mis)information ($M = 9.47, SD = 0.08$) than retractions ($M = 9.10, SD = 0.24$), $r(317) = 12.44, p = 3.36e - 29$.

Next, we normalized belief ratings and activations for each participant or model run by dividing each rating/activation for (mis)information and retractions by their sum. Figure 3 shows the normalized values for (mis)information and retractions across all six conditions for human data and model. We assessed the model fit to human data with the same procedure used for critical (mis)information scores. The model had a negative non-significant relationship with the trends in the human data, $r(10) = -0.29, p = 0.57702$, and a reasonable average difference overall, $RMSE = 0.11$.

Discussion

We presented an initial model to explore memory processes underlying the CIE. Our previous demonstration (Hough et al., 2023) supported notions in the literature, where the CIE occurs because corrective information competes with, rather than overwrites, well-established memories. Our current ef-
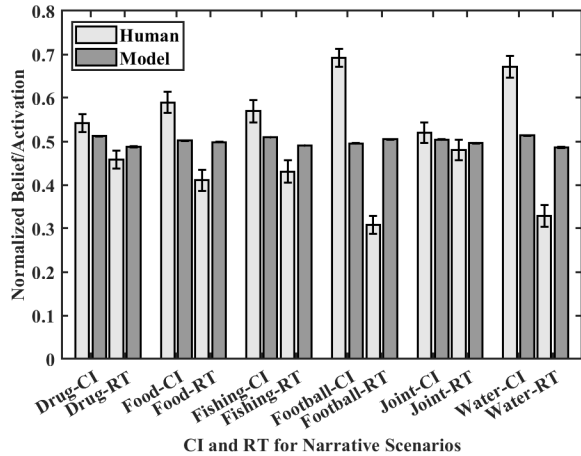
Figure 3: Normalized belief ratings from human data and model chunk activations for critical (mis)information (CI) and retractions (RT) for all six scenarios. Error bars are SEM

fort demonstrates how misinformation and retractions compete based on how interconnected they are within the scenario in a context without learning opportunities. In the model, the competition involves the activation of chunks in memory representing the critical (mis)information and retractions.

Processing and producing language in a cognitively plausible way is a critical challenge in cognitive modeling. To compare the model with human data, we explored methods to approximate behaviors like text processing, question answering, and beliefs. We made progress on all three and were able to present model fits to critical (mis)information and belief scores. We had to reorganize the human data by collapsing source manipulation conditions and extracting the data for scenarios. Critical (mis)information scores were rather difficult to approximate. We had to further modify the human data and create a rather complicated method to produce scores for the model. The resulting model fit to critical (mis)information scores was interesting, but weak. Despite the difficulty and weak fit, we felt it was important to include this measure and comparisons as it is commomly used in the literature. For belief scores, the methods were more straightforward. The human data did not require modification, and we were able to approximate beliefs for the model by calculating the average activation for chunks associated with the critical (mis)information and retractions. The fit to belief scores were promising and provided a more clear method to use in the future.

Limitations and Future Work

We note several limitations with our current work: 1) We manually generated predicates and focused on declarative memory with chunk chaining, 2) we selected questions from Ecker and Antonio (2021) congruent to the reasoning and expression capacities of our model (i.e while inferential reasoning was too complex for our model, belief rating was possi-

ble), 3) we had to do some irregular modifications to compare the model and human data, 4) we left out two components from our targeted gap: social and emotional factors, and 5) we did not include mechanisms to access source information.

Language processing is an important component for models focused on cognitive vulnerabilities and biases in the information environment. Here, we parsed the scenarios using ChatGPT, but its output method changed and we had to manually generate predicates for five of the six scenarios. The parsing and representation of written content has a large impact on a model's behavior, and how a model interprets questions constrains the types of responses that are possible. For instance, the current model will be heavily influenced by which information can be chained together. It is unclear how to best handle these issues, which will be an important issue for future work.

Given the difficulties with having the model answer questions from (Ecker & Antonio, 2021), we had to approximate model responses. This created some difficulties with comparing to human behavior. However, we learned that chunk activations are a straightforward method to assess the model's belief or bias towards types of information.

We targeted the research gap highlighting our lack of understanding of the interplay between cognitive, social, and emotion factors. However, we only focused on cognition, specifically, memory. We plan to include these factors in the future, starting with emotion. We plan to implement emotion mechanisms using a previously developed ACT-R module (Juvina et al., 2018) based on core-affect theory (Russell & Barrett, 1999) that focuses on valence (i.e., positive or negative) and arousal (i.e., intensity). However, we are unaware of a study exploring the CIE and manipulating emotion to assess affects on memory. Therefore, we plan to conduct some experimental work needed to validate a CIE model that includes emotion.

Lastly, we did not include mechanisms to interpret source trust and credibility. However, our planned future work with emotion and core affect has the potential to inform these mechanisms. For instance the core-affect module was used in previous work (Juvina et al., 2019) to model trust and therefore, can serve as or inform mechanisms to model all the conditions from Ecker and Antonio (2021).

Conclusion

Overall, we demonstrated a CIE across scenarios using only basic components of declarative memory and memory navigation. These general mechanisms and processes are not specific to the CIE and can be used for any information, given it is presented in word pairs. It was clear that the knowledge representations significantly impact how the model navigates memories for scenarios. This needs to be considered when knowledge is engineered for models and in future experiments to control for a potential confound where misinformation is more interrelated in the narrative than corrections. Our modeling approach and results aligned well with the mental model explanation in the literature and with availability of in-

formation in memory or ease of recall. The model provides a good base to extend in future research by adding emotion and social factors and testing theoretical explanations across experiments and datasets.

Acknowledgments

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195324259.001.0001

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin and Review*, *5*, 1–21. doi: 10.3758/BF03209454

Brydges, C. R., Gignac, G. E., & Ecker, U. K. (2018). Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. *Intelligence*, *69*, 117–122.

Ecker, U. K., & Antonio, L. M. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, *49*, 631–644. doi: 10.3758/s13421-020-01129-y

Ecker, U. K., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of applied research in memory and cognition*, *6*(2), 185–192.

Ecker, U. K., Lewandowsky, S., Cheung, C. S., & Maybery, M. T. (2015). He did it! she did it! no, she did not! multiple causal explanations and the continued influence of misinformation. *Journal of memory and language*, *85*, 101–115. doi: 10.1016/j.jml.2015.09.002

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*, 570–578. doi: 10.3758/s13423-011-0065-1

Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*, 1087–1100. doi: 10.3758/MC.38.8.1087

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending sme to handle large-scale cognitive modeling. *Cognitive Science*, *41*(5), 1152–1201.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601–613. doi: 10.1037/0022-3514.59.4.601

Halverson, T., Stevens, C., Fisher, C., Haubert, A., & Myers, C. (2018). Balancing confidence and information costs in a diagnostic reasoning task. In *Proceedings of the 16th international conference on cognitive modeling.* Madison, WI: University of Wisconsin.

Hough, A. R., Fisher, C., Stevens, C., Curley, T., Larue, O., & Myers, C. (2023). Modeling the continued influence effect in the information environment. In *16th sbp-brims conference.*

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436. doi: 10.1037/0278-7393.20.6.1420

Johnson, H. M., & Seifert, C. M. (1998). Updating accounts following a correction of misinformation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1483–1494. doi: 10.1037/0278-7393.24.6.1483

Juvina, I., Collins, M. G., Larue, O., Kennedy, W. G., Visser, E. D., & Melo, C. D. (2019). Toward a unified theory of learned trust in interpersonal and human-machine interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *9*(4), 1–33.

Juvina, I., Larue, O., & Hough, A. (2018). Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. *Cognitive Systems Research*, *48*, 4–24. doi: 10.1016/j.cogsys.2017.06.002

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. doi: 10.1177/1529100612451018

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433–449. doi: 10.1016/j.jesp.2003.07.008

Myers, C. W., Gluck, K. A., Harris, J., Veksler, V., Mielke, T., & Boyd, R. (2015). Evaluating instance-based learning in multi-cue diagnosis. In *Na taatgen, mk, van vugt, j. p borst, & k. mehlhorn. proceedings of iccm 2015. paper presented at international conference on cognitive science, groningen, netherlands (198-199).*

Romero, O. J., Zimmerman, J., Steinfeld, A., & Tomasic, A. (2023). Synergistic integration of large language models and cognitive architectures for robust ai: An exploratory analysis. In *Proceedings of the aaai symposium series* (Vol. 2, pp. 396–405).

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, *76*(5), 805. doi: 10.1037/0022-3514.76.5.805

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? In *Psychology of learning and motivation* (Vol. 41, pp. 265–292). Elsevier. doi: 10.1016/S0079-7421(02)80009-3

Swire, B., Ecker, U. K., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1948–1961. doi: 10.1037/xlm0000422

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, *240*, 112552. doi: 10.1016/j.socscimed.2019.112552

Wilkes, A., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology*, *40*(2), 361–387. doi: 10.1080/02724988843000168

Understanding Emotion and Emotional Contagion Effects
on Cooperative Behavior through Game Simulation

Ruiki Kawaji (kawaji.ruiki.16@shizuoka.ac.jp),
Junya Morita (j-morita@inf.shizuoka.ac.jp)
Department of Informatics, Graduate School of Integrated Science and Technology, Shizuoka University,
3-5-1, Johoku, Hamamatsu Chuo-ku, Shizuoka, 432-8011, Japan

Hirotaka Osawa (osawa@a3.keio.jp)
Department of Industrial and Systems Engineering, Faculty of Science and Technology, Keio University,
3-14-1, Hiyoshi, Yokohama Kohoku-ku, Hamamatsu, Kanagawa, 223-8522, Japan

Abstract

In this study, we considered the influence of emotion and its contagion on the success of cooperative behavior. We conducted a simulation using the board game "Hanabi", which is suitable for analyzing cooperative behavior. The results of the simulation showed a decrease in the score of the model with emotion, confirming the negative influence of emotion on cooperative behavior. The analysis indicated that cooperative behavior was more successful when valence and arousal were higher. Furthermore, it was suggested that arousal level was more likely to induce synchronization through emotional contagion than emotional valence.

Keywords: ACT-R, Cognitive modeling, Cooperative behavior, Emotion, Hanabi

Introduction

In contemporary society, group psychology wields significant influence, as a few emotional opinions can often mobilize a large group of people who align with them. This phenomenon is termed group polarization (Stoner, 1968). When group polarization occurs, individuals may undertake risks collectively that they would not consider individually, or conversely, they may adopt excessively cautious approaches. Several studies on risk shifting have explored the relationship between emotion and risk shifting (Turner, Zangeneh, & Littman-Sharp, 2006; Yuen & Lee, 2003). Just as emotions are linked to individual risk shifting, the polarization of risk within groups may stem from emotional contagion. According to evolutionary psychology (Cosmides & Tooby, 1997), emotions evolve due to their utility in survival and their facilitation of cooperation and socialization. In this context, group polarization appears instrumental in fostering group consensus. However, this does not necessarily imply its usefulness in arriving at rational conclusions. Resolving this question necessitates clarifying the role and utility of emotions within a group.

Numerous studies have explored emotions, proposing various models to conceptualize them. For instance, Russell's circumplex model describes emotions along two axes (Russell, 1980), while Barrett's model constructs emotions based on predictions (Barrett, 2017). This study advocates for implementing these models within computer simulations to objectively validate their efficacy. Our model is developed base on a cognitive architecture, Adaptive Control of Thought-Rational (ACT-R: Anderson, 2007). Employing a uniform architecture such as ACT-R for these simulations enables comparison and integration of these models.

This study utilizes the cooperative board game "Hanabi" as a task to observe the impact of emotional models on behavior. Through this investigation, we aim to underscore the significance of emotions in facilitating cooperative behavior. Additionally, our findings confirm the occurrence of emotional contagion among models communicating through the environment. Concurrently, we develop a composite model integrating physiological and psychological emotion models within existing cognitive architectures, thereby reaffirming the utility of simulation in evaluating diverse emotion models.

Related Research

Emotion and risk shifting

Two competing theories regarding risk shifting exist. One posits that individuals are more inclined to take risks following successes (Thaler & Johnson, 1990), while the other suggests they are less inclined after successes and more so after failures (Leopard, 1978). Turner et al. (2006) identified issues with experiments conducted in previous studies and thus conducted a new experiment to address them. The results indicated that participants who experienced success were more inclined to gamble recklessly compared to those who experienced failure when gambling again. Questionnaire analysis revealed a moderate predictive power for emotions and risk shifting.

Studies employing direct mood induction to measure changes in risk shifting have shown mixed results. Specifically, they have not consistently contrasted risk shifting under positive versus negative emotional states (Yuen & Lee, 2003). The findings of this study demonstrated a relatively weak tendency to accept risk when experiencing positive emotions and a strong tendency to avoid risk when experiencing negative emotions. The reasons for these findings were further discussed in terms of variations in judgment methods based on participants' emotional states.

The mood congruence effect elucidates the correlation between emotions and risk shifting (Bower, 1981). It is a phenomenon wherein memories are more readily recalled when the current mood aligns with the mood of the mem-

ory. This implies that positive moods tend to evoke recollection of successful experiences, leading individuals to misperceive the current situation and take risky actions based on past successes. Conversely, negative moods may evoke memories of failures, prompting individuals to adopt a risk-averse stance. Conversely, the mood incongruence effect serves as a counterpart to the mood congruence effect (Parrott & Sabini, 1990). In this scenario, individuals tend to recall failures when in a positive mood and successes when in a negative mood. Experiments conducted by Rinck, Glowalla, and Schneider (1992) suggest that the mood incongruence effect manifests partially when emotions are weak, with the mood congruence effect predominating.

Emotion and Recall with ACT-R

ACT-R is a cognitive architecture developed based on the ACT-R (Adaptive Control of Thought-Rational) theory proposed by Anderson and Bower (2014). This theory outlines the functions and structures of the human brain, and ACT-R is designed to model and simulate these functions and structures accordingly. It operates as a production system, processing information using rules and executing actions based on two forms of knowledge: declarative knowledge, which describes facts, and procedural knowledge, which manages methods such as actions. A notable advantage of a production system is its ability to handle information symbolically. ACT-R also has strength of dealing information that may be difficult to represent as symbols by using subsymbolic processing.

Various studies have utilized ACT-R to model and express emotions. For instance, Sakai, Itabashi, and Morita (2022) constructed a model aimed at presenting the user's emotions based on the estimated emotional state, utilizing the recall method commonly used in mental health care. In their model, emotions were represented by two parameters derived from Russell's circumplex model, and recall operations were conducted in accordance with the user's emotional state by mapping them to the parameters within ACT-R. The study demonstrated the influence of emotion on recall and behavioral changes by associating valence with utility values and arousal with noise values in the ACT-R framework.

As a means of incorporating emotional influence into computer models, Juvina, Larue, and Hough (2018) integrated an emotion module into the ACT-R cognitive architecture. Their simulation results were consistent with experiments on mood congruence effects, indicating that memory recall could be modeled to reflect human emotions. Emotions in the study were represented by two parameters based on the dimensionality theory of emotion: Valuation and Arousal. These parameters were dynamically adjusted based on memory evaluation and added to the activation used in ACT-R calculations to represent the emotional effect.

Hanabi

Hanabi is a cooperative board game that offers a rich environment for analysis and modeling due to its inherent re-

quirement for cooperation and well-defined behavioral dynamics, including communication, within the game. There has been many studies developing cooperative agents that play this game (Bard et al., 2020; Bouzy, 2017). As an example, Osawa, Kawagoe, Sato, and Kato (2021) developed models of Hanabi and its players to investigate various aspects of human behavior, such as self-estimation based on the estimation of others. Additionally, Kawagoe and Osawa (2022) utilized ACT-R to examine the effects of differing risk shifting between players on reaction time and cooperative behavior. Moreover, Kuwabara et al. (2023) explored cooperative behavior through intention estimation, employing a model that leverages the similarity between players' hands and the situational context during instance retrieval. Building upon the insights from these studies, we aim to model the influence of emotion on the success of cooperative behavior.

The Rules of Hanabi

This study aims to demonstrate that cooperative behavior can be effectively guided by emotions, through simulations with Hanabi as a task. To facilitate this investigation, we first outline the rules of Hanabi.

Hanabi is designed for two to five players; however, for the sake of simplicity in modeling cooperation, this study focuses solely on two-player games. The game comprises 50 cards in five colors and two types of tokens. Each color (white, red, blue, yellow, and green) consists of 10 cards, including three 1's, two 2's to 4's, and one 5. The objective of the game is to collaboratively assemble five fireworks displays by stacking cards of the same color in numerical sequence and placing a 5 card of each color on the field.

At the onset of the game, the cards are shuffled, and five cards are distributed to each player. Players cannot view the cards in their own hands; however, these cards are visible to the other players. The remaining cards form the deck. The first player begins their turn with eight blue tokens and zero red tokens, which are collectively shared among all players.

Players take turns, and the next player's turn commences after the current player performs one of the following three actions:

1. Hint: The player selects either a color or a number of a card that is not their own and reveals which cards in their hand correspond to the selected information. If multiple cards match the selected information, all corresponding cards must be indicated. Providing a clue incurs a cost of one blue token, with each clue consuming one token. If there are no blue tokens available to cover the cost, this action cannot be performed.

2. Discard: The player discards a card from their hand and draws a new card from the deck. The new card is visible only to the opponent. The discarded card is revealed to all players and cannot be reused during the game. Discarding a card recovers one blue token. However, the player cannot accumulate more blue tokens than the initial value.

3. Play: The player reveals a card to the public and determines whether the play is successful. If the card's value is exactly one more than the current top card of the corresponding color in the fireworks display, the play succeeds, and the card is placed atop the fireworks display. If there is no card of the same color in the display, it is considered as 0, and the card can be placed on top. Otherwise, the play fails, and the card is treated as discarded. In both cases, a card is drawn from the deck, visible only to the opponent. Successfully playing a 5 card and completing a fireworks display of one color earns a bonus blue token.

The game concludes under one of three conditions: accumulating three red tokens, all players taking one final turn after the deck is depleted, or completing all five colored fireworks displays. Upon meeting the end condition, the score is calculated as the sum of the cards stacked in the fireworks display. The maximum score is 25 points (5 colors × 5 cards).

Model

In order to capture changes in behavior influenced by emotion, an agent model was developed using ACT-R. Given its utilization in previous studies for determining behavior and measuring emotional effects on behavior, ACT-R was deemed appropriate for constructing the model in this study. Figure 1 depicts a flowchart illustrating the process of the model employed in this study. It builds upon the previous model (Kuwabara et al., 2023), incorporating adjustments tailored for emotion-adapted behavior.
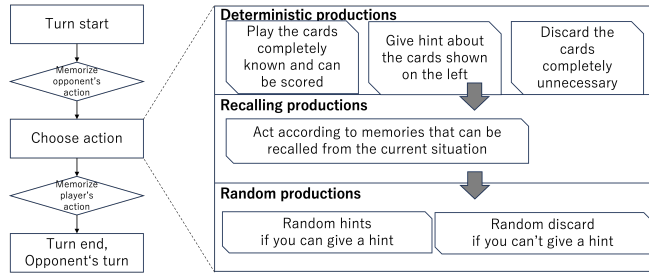

Figure 1: Flowchart of the model

Accumulation of Instance

As illustrated in the flowchart, our model accumulates and utilizes examples. Before and after each agent executes an action in the game, it assesses whether its own or its opponent's preceding action was successful and memorizes this information if no identical memory exists. The information encapsulated in the instances comprises the following:

- Success or failure of the action
- Type of action
- Information of the target card
- The state of the field
- The player's visible hand

- The opponent's visible hand
- The all of opponent's hand
- Information on the remaining cards

"Types of actions" are limited to three: playing cards, giving hints, and discarding cards. When playing or discarding cards, the "information of the target card" encompasses its color and number. In the case of giving hints, the "information of the target card" may pertain to either the color or the number of the card. If the color or number of a card is not yet known when playing or discarding, it is recorded as "unknown."

The determination of "success or failure of the action" varies based on the "types of action." For hinting, success is achieved if the opponent takes immediate action following the hint. For discarding a card, success is determined if the discarded card is not the last card of the same color and number.

"The state of the field" denotes the quantities of cards of each color (white, red, blue, yellow, and green) present on the field. "The player's visible hand" comprises cards known to the player through clues provided by the opponent. "The opponent's visible hand" comprises cards known to the player through clues provided by the opponent. The complete information about the opponent's hand, "the all of opponent's hand," is not visible to the opponent.

"Information of the remaining cards" indicates the total number of cards in Hanabi minus the number of confirmed cards. Confirmed cards include those that have been played, discarded, or are in the opponent's hand, as well as cards in the player's hand whose color and number are known to the player.

Instance-based Learning

When deterministic action is not feasible, as depicted in the flowchart, the model resorts to decision-making via recall. The initial step involves searching for past instances based on the current situation. The ACT-R function, partial matching, is utilized to retrieve instances, with the closest matches recalled, even if not identical. If the recalled instance denotes success, the agent endeavors to act accordingly. However, if the recalled instance signifies failure, it is recalled once more to avoid repeating the same action type. Given the presence of three action types, the maximum number of recalls is three. In instances where no recall occurs or if the agent fails to adhere to the recalled instance, a random decision occurs.

In the partial matching mechanism of ACT-R, the activation A_i of each memory i is computed using the following equation, with the memory possessing the highest activation recalled:

$$A_i = B_i + \sum_l PM_{li} + \varepsilon \qquad (1)$$

B_i is the base level, ε is the noise, and P is the discrepancy penalty coefficient for partial matches. In our model, each chunk in the instance is weighted equally. Consequently, the

importance of a memory element (chunk) l is represented as $T_l(\sum_l T_l = 1)$, and the activation is expressed as follows:

$$A_i = B_i + \sum_l PT_l M_{li} + \varepsilon \qquad (2)$$

Table 1: Table of importance

Memory element l	importance T_l
Success or failure of the action	0.5
The state of the field	0.2
Colors of player's visible hand	0.06
Numbers of player's visible hand	0.06
Colors of opponent's visible hand	0.06
Numbers of opponent's visible hand	0.06
Colors of all of opponent's hand	0.025
Numbers of opponent's visible hand	0.025
Information on the remaining cards	0.01

As illustrated in Table 1, T_l is assigned a large value for information that is commonly available to both the player and the opponent, and is deemed valuable for the most recent decision-making process. The calculation of M_{li}, the similarity, is conducted through three different methods based on various factors. For the state of the field and information on the remaining cards, it is computed as (number of matched elements/number of elements)-1. The number of elements considered is 5 (representing the number of colors) for the state of the field, and 25 (5 colors $\times$5 numbers) for the information on the remaining cards. Regarding information on the cards in the hand, the data is categorized into colors and numbers, then transformed into a frequency vector, with the cosine similarity utilized to determine similarity, where the cosine similarity of -1 denotes M_{li}. However, in the instance where the information is represented as 0 (indicating unknown), we didn't calculate it as such, but instead added 0.2 to each of the five values to represent unknown. For determining the similarity based on the success or failure of an action, the difference between the current valence V and the success or failure of the target instance V_i s calculated as follows:

$$M_{li} = -|V - V_i|/2 \qquad (3)$$

Here, V_i equals 1 for a success instance and -1 for a failure instance.

Emotional Influence

The model's emotion is characterized by two parameters: Valence and Arousal, which influence the recall process. Valence serves as a memory component during similarity calculations, reflecting the mood congruence effect, a phenomenon where emotional valence affects recall. The emotional distance between the remembered instances is assessed by labeling them as positive if successful and negative if unsuccessful, with closer instances being more likely to be recalled.

Arousal level modulates the value of BLC, an offset parameter of base-level, ranging from 0.5 to 1.5 times. This adjustment affects the base level B_i, as mentioned previously, making successful recall more likely under high arousal and more prone to failure under low arousal.

Update of emotion in the model follows the concept of prediction error. Table 2 summarizes our policy of changing the emotional parameters. Those policies are based on the model of Joffily and Coricelli (2013), where prediction error is pleasant when it converges towards resolution and unpleasant when it diverges. To implement this in our model, we varied emotion fluctuations based on the types of acts (deterministic, recalling, or random acts). In deterministic acts (always success in evaluation with no recall), emotions remain relatively stable since they consistently succeed, resulting in minimal prediction-result disparity. In random acts, where both prediction and experience accuracy are low due to action uncertainty, emotions exhibit minimal change (randomly success or failure in evaluation with no recall). However, in recalling acts, the model anticipates success (with recall), leading to significant emotional change based on prediction error resolution. If successful, emotions move towards positive, while failure leads to negative emotion. Recalling behavior triggers substantial arousal changes but without recall, slightly decreases arousal over time.

Table 2: Act and Emotional Changes

Evaluation	Recall	Valence	Arousal
Success	Yes	Increase large	Increase large
Success	No	Increase small	Decrease small
Failure	Yes	Decrease large	Increase large
Failure	No	Decrease small	Decrease small

More specifically, the emotion is updated as follows:

$$V_{t+1} = V_t + \alpha(r - V_t) \qquad (4)$$

For valence variability, $r = 1$ for success and $r = -1$ for failure, $\alpha = 0.2$ for recall use and $\alpha = 0.02$ for non-use. For arousal variability, when recalling $r = 1$ and $\alpha = 0.2$ for recall, while for non-use, $r = -1$, $\alpha = 0.01$.

Simulation

We conducted simulations using two agents of the model described in the previous section to assess the impact of emotion on cooperative behavior. Additionally, for comparison, we conducted simulations using agents of the model without emotion to contrast from the model with emotion.

Simulation method

In the simulation, we conducted 100 consecutive trials, repeated 10 runs, where each trial represented a complete game from start to finish. The average of the 10 runs across the 100 trials was calculated to observe learning through repeated trials.

Both Valence and Arousal were reset to 0 at the onset of each trial while memorized instances persisted for the duration of one run. Memorized instances were reset at the initiation of each run. To eliminate bias between preceding and following players, Agent 1 initiated odd-numbered trials, while Agent 2 initiated even-numbered trials.

For comparative analysis, a simulation of the model without emotion was conducted with the above same condition. Agents without emotion set at a fixed ± 0 throughout the simulation.

Results

Scores with and without Emotion   The average score of 10 runs per trial was computed for both simulations with/without emotions. Figure 2 presents the results. Additionally, for comparison, we included the average score from a previous study (Kuwabara et al., 2023) that also utilized past instances to determine behavior.
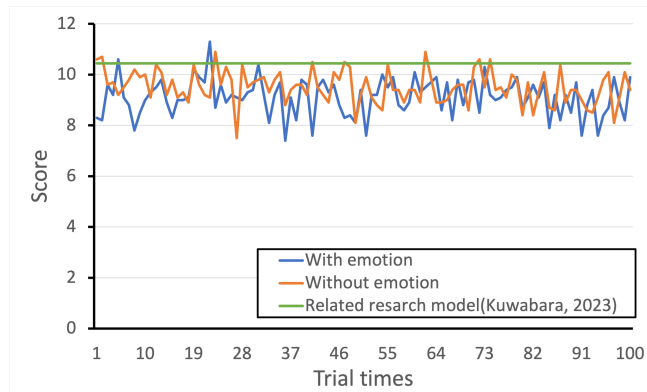

Figure 2: Simulation results

The graph illustrates that the outcomes of this study are lower compared to those of the previous study. Specifically, the score of the model with fluctuating emotions ($mean = 9.118$, $SD = 2.034$, $n = 1,000$) was significantly lower ($t(1998) = 4.433$, $p < 0.001$, $d = 0.198$) than that of the model without fluctuating emotions ($mean = 9.516$, $SD = 1.981$, $n = 1,000$).

The lower score in the present model compared to the previous study could potentially stem from the reliance on failure instances for search. This hypothesis aligns with the finding that the model with emotion yielded a lower score than the model without emotion. When emotional fluctuations lean towards negativity, likelihood of recalling failure experiences increase, leading to excessive conservative cautious sift. Consequently, the probability of successful behavior diminishes due to the random behavior induced by emotional fluctuations.

Furthermore, we investigated how the emotional effects mentioned above were influenced by the accumulation of trials. In the model without emotions, there was a significant decreasing score trend ($r = -0.251$, $p < .01$) along with the

number of trials. However, in the model with emotion, although the correlation between the number of trials and the score was negative ($r = -0.068$, $p = 0.498$), it did not reach statistical significance.

This outcome suggests the negative effect of an accumulation of failed instances. Examination of the execution logs revealed that over 99% of the models' executed actions that were classified as failures when providing hints. Several factors may contribute to hints being classified as failures. Firstly, players don't always proceed with gameplay immediately after receiving a hint that thoroughly clarifies the game, resulting in a failure judgment. Secondly, in scenarios where numerous crucial hints are available, deterministic actions are prioritized for providing hints, while hints utilizing recall are more commonly utilized for less critical hints. This tendency may stem from accumulated experiences with unsuccessful hints across various situations, facilitating easier recall of such unsuccessful hints and subsequently leading to their avoidance.

The cautious cooperative behavior observed in the simulations is thought to be linked to cautious shift, a phenomenon within group polarization. Throughout the simulation, we frequently encountered instances where random behavior persisted following a random action, stalling the progression of the situation. In the current model, random actions were executed when deterministic actions weren't feasible and successful instances weren't recalled. The failure to recall successful instances with a sufficient number of trials is indicative of a scenario where emotional valence is negative, leading to the recall of only failure instances. Consequently, the model sought to avoid risk, resulting in the execution of random actions. This cascade of risk-averse behavior influencing the partner's risk aversion, and the subsequent chain reaction of risk-averse behavior, is regarded as a manifestation of the cautious shift phenomenon.
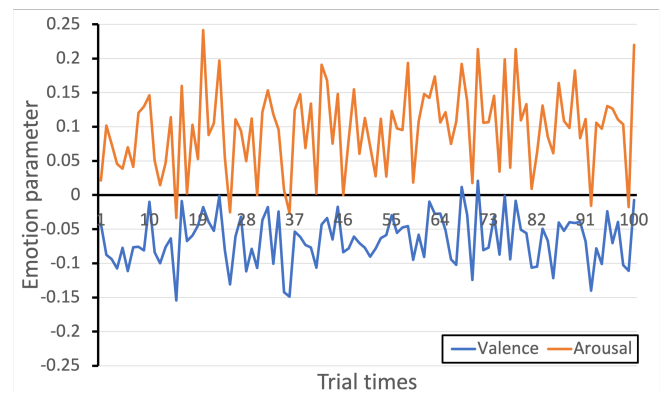

Figure 3: Emotional Parameters

Emotional Changes   To analyze emotional changes during the simulation, we calculated the average of emotional parameters for the each trial. The results are depicted in Figure 3. It's evident from the figure that valence tended to be nega-

tive overall ($mean = -0.070$, $SD = 0.163$, $n = 1000$, one-step t value against 0 $t(999) = -13.480$, $p < .01$) while arousal tended to be positive ($mean = 0.090$, $SD = 0.210$, $n = 1000$, one-step t value against 0 $t(999) = 13.514$, $p < .01$). This finding aligns with the preceding section, indicating that the recall of failure instances induced risk-avoiding behavior, consequently resulting in lower scores compared to the model without emotion.

Regarding the correlation between the trials and the emotion parameters aggregated across model runs, no significant correlations were observed (valence: $r = 0.031$, $n = 100$, arousal: $r = 0.066$, $n = 100$). However, to explore the relationship between emotions and scores, correlation coefficients were calculated between the emotion parameters and scores aggreated across two agents. The results revealed a significant correlation for valence ($r = 0.370$, $n = 1000$), indicating that higher emotional valence corresponds to higher scores, and a significan positive correlation for arousal ($r = 0.237$, $n = 1000$), suggesting that higher arousal levels are associated with higher scores within the scope of this simulation.

Emotional Contagion   To investigate whether emotional contagion occurs alongside individual emotion fluctuation, correlation coefficients between agents were computed for each of the 10 runs (Table 33). The maximum correlation coefficient for arousal was 0.662, while the minimum was 0.486, signifying a positive correlation while all the correlation for valence are low values.

Table 3: Emotion correlations between agents

Executions	Valence	Arousal
1	-0.033	0.607
2	-0.011	0.493
3	0.160	0.550
4	0.006	0.496
5	0.152	0.535
6	0.071	0.662
7	-0.007	0.557
8	-0.005	0.549
9	0.011	0.506
10	0.062	0.486

This suggests that emotional valence, such as positivity or negativity, doesn't exhibit strong contagion in our model. However, arousal level appears to be adequately contagious through the environment. The discrepancy in the contagion of the two emotional parameters warrants further consideration.

Summary and Future Works

In this study, we developed and simulated a two-agent cooperative game model to explore the interplay between successful cooperative behavior and emotion. Our model adjusts emotion based on disparities between the environment and predictions, thereby influencing behavior through emotional recall. The findings revealed that the inclusion of emotion led to a decrease in the overall score compared to the model without emotions, providing evidence of the detrimental impact of emotion on cooperative behavior.

In the simulation model utilized in this study, it was observed that negative valence leads to excessive risk avoidance, ultimately resulting in a lower overall score. These findings suggest a potential association with the cautious shift phenomenon, as a cascade of risk-avoiding behaviors.

The findings indicate that emotional contagion varies between two parameters based on the dimensionality theory of emotion: valence and arousal. Simulations conducted using the model in this study revealed that arousal exhibits a stronger correlation compared to valence, suggesting that contagion is more likely to occur based on arousal levels.

While emotions did not increase the scores, we believe that there is room for improvement through parameter adjustments. Despite observing risk-avoiding behavior in the model, there is potential to foster a more risk-taking approach by fine-tuning the parameters. Application of this model to human experiments holds promise for gaining deeper insights into agent interactions, particularly with varied adjustments accounting for individual differences.

In the scope of our simulation results, positive values of the emotion parameter exhibited a beneficial impact on cooperative behavior. However, the range of the varied emotion parameter was limited, thus hindering the assurance of generality. Therefore, we posit that conducting simulations where the emotion parameter is fixed at various values other than 0 would provide insights into the effects of specific emotional states and the utility of fluctuating emotions. We anticipate that such comparisons will enhance our understanding of how different emotional states and their variability influence behavior.

There is also room for improvement in the structure of the model itself. While in this study, the influence of emotion was primarily manifested in the change in the activation, other studies exploring behavioral changes due to emotion suggest that strategies may vary depending on emotional states. Many other features of ACT-R, such as spreading activation and blending, can be incorporated in our model. We believe that incorporating these functions could enhance the model's validity.

Compared to previous studies utilizing the same ACT-R framework for simulating Hanabi, the simulation time in our study increased due to heightened computational complexity, leading to the lack of simulation runs. The model necessitates memorizing both successful and unsuccessful instances, and multiple recalls for a single action, prompting a need to devise methods to mitigate computational complexity.

References

Anderson, J. R., & Bower, G. H. (2014). *Human associative*

memory. Psychology press.

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., ... others (2020). The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, *280*, 103216.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Bouzy, B. (2017). Playing hanabi near-optimally. In *Advances in computer games: 15th international conferences, acg 2017, leiden, the netherlands, july 3–5, 2017, revised selected papers 15* (pp. 51–62).

Bower, G. H. (1981). Mood and memory. *American psychologist*, *36*(2), 129.

Cosmides, L., & Tooby, J. (1997). *Evolutionary psychology: A primer* (Vol. 13). Citeseer.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS computational biology*, *9*(6), e1003094.

Juvina, I., Larue, O., & Hough, A. (2018). Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. *Cognitive Systems Research*, *48*, 4–24.

Kawagoe, A., & Osawa, H. (2022). The proposal of cooperative behavior analysis using cognitive model in hanabi game. In *The 36th annual conference of the japanese society for artificial intelligence(2022)* (pp. 4I3OS26b04–4I3OS26b04).

Kuwabara, R., Nagashima, K., Morita, J., Miyata, K., Kawagoe, A., & Kawagoe, H. (2023, may). Constructing a communication model for inference using the cooperative game hanabi. *Human-Agent Interaction Symposium 2023*.

Leopard, A. (1978). Risk preference in consecutive gambling. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 521.

Osawa, H., Kawagoe, A., Sato, E., & Kato, T. (2021). Emergence of cooperative impression with self-estimation, thinking time, and concordance of risk sensitivity in playing hanabi. *Frontiers in Robotics and AI*, *8*, 658348.

Parrott, W. G., & Sabini, J. (1990). Mood and memory under natural conditions: Evidence for mood incongruent recall. *Journal of personality and Social Psychology*, *59*(2), 321.

Rinck, M., Glowalla, U., & Schneider, K. (1992). Mood-congruent and mood-incongruent learning. *Memory & cognition*, *20*(1), 29–39.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sakai, S., Itabashi, K., & Morita, J. (2022). Estimating personal model parameters from utterances in model-based reminiscence. In *2022 10th international conference on affective computing and intelligent interaction (acii)* (pp. 1–8).

Stoner, J. A. (1968). Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology*, *4*(4), 442–459.

Thaler, R. H., & Johnson, E. J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management science*, *36*(6), 643–660.

Turner, N. E., Zangeneh, M., & Littman-Sharp, N. (2006). The experience of gambling and its role in problem gambling. *International Gambling Studies*, *6*(2), 237–266.

Yuen, K. S., & Lee, T. M. (2003). Could mood state affect risk-taking decisions? *Journal of affective disorders*, *75*(1), 11–18.

Memory Activation and Retrieval Strategy in Lexical Alignment: Comparing the ACT–R Model of Human and Computer Interlocutors

Miki Matsumuro (mm3398@cornell.edu)
Cornell University Department of Communication, Mann Library, 450
237 Mann Dr, Ithaca, NY 14853 USA

Yugo Hayashi (yhayashi@fc.ritsumei.ac.jp)
Ritsumeikan University, College of Comprehensive Psychology, 2-150 Iwakura-cho
Ibaraki, Osaka 567-8570 JAPAN

Abstract

During conversations, speakers tend to reuse the lexical expressions of their interlocutors. This is called "lexical alignment," and it facilitates the listener's understanding of the speaker's intention. Branigan et al. (2011) has shown that this tendency increases when speakers believe that their partner is a computer agent rather than a human. Memory activation for the expressions used by the interlocutors and the strategy preference whereby speakers attempt to use their partners' expressions rather than those that first come to mind have been shown to be the causes of lexical alignment. For this study, we constructed an ACT–R model for which we could adjust the parameter values related to these two features. Through parameter adjustment, we simulated lexical alignment with both human and computer agents in Branigan et al. (2011). For both partner conditions, additional activation was added to the knowledge of the partners' expressions. The computer–partner model preferred trying to retrieve the partners' expression rather than using the knowledge that had a strong association with the stimulus and was easy to retrieve. In contrast, the human–partner model had no specific preference; that is, it displayed equal utility for both. A comparison of these parameter values revealed that the computer–partner model preferred to retrieve the partner's knowledge; in addition, it also kept the knowledge's activation sufficiently high so that it could be available for a longer duration.

Keywords: Lexical alignment; ACT–R; Human–Computer Interaction; Cognitive model

Introduction

Lexical alignment is a behavior observed during conversations that involves a speaker reusing their interlocutor's lexical expressions. It helps the speaker communicate smoothly, and its usage increases when users interact with computer agents. In this study, we constructed cognitive models and investigated whether memory activation and the frequency with which the speaker attempts to retrieve the interlocutor's expressions—both of which are considered causes of lexical alignment—explain the occurrence of lexical alignment. The model's parameter adjustment reveals the source of the difference in lexical alignment between human and computer partners.

Factors Influencing Lexical Alignment

When talking with a partner, there is an expectation that the listener correctly understands the intentions of the utterance. Speakers use various strategies to facilitate listeners' understanding (e.g., Clark & Murphy, 1982; Pickering & Garrod, 2004), and lexical alignment is one such strategy (Brennan &

Clark, 1996; Garrod & Anderson, 1987). People tend to reuse lexical expressions that their interlocutor used previously.

Memory Retrieval Memory activation is a process involved in lexical selection. As shown in priming studies, recently used words maintain a higher activation and tend to be selected (e.g., Meyer & Schvaneveldt, 1971). Syntactic knowledge is also primed by preceding sentences (Pickering & Branigan, 1998). These studies suggest that lexical alignment occurs because a lexical expression is activated when it is used by the interlocutor, facilitating easy retrieval.

Top-Down Processes Speakers carefully decide which lexical expression is the most appropriate for ensuring that listeners can easily and correctly understand their intentions. For instance, when two Japanese people are speaking with each other, the sentence "I live in Shiga" is sufficient to indicate to the interlocutor where the speaker lives. However, when speaking to an interlocutor who is not from Japan, a broader description, such as "I live in the prefecture next to Kyoto," would be necessary to provide a more accurate understanding.

As described in the previous example, the speaking partner's attributes (e.g., nationality, age, gender, etc.) are an important factor when deciding which phrase should be used (Fussell & Krauss, 1992). These attributes are defined as top-down factors because speakers use knowledge-based strategies to decide which phrase should be used in a particular context.

Communication with Computer Agents

With the rapid spread of ChatGPT, communication with computer agents has become more commonplace, even among those not involved in higher education. Researchers in the field of human–agent interaction examine the aspects of agents that impact the communication process, such as visual appearance (Baylor, 2009). Several studies have shown that when a speaker believes that their partner is a computer, they communicate with them differently than when they believe their partner is a human (Amalberti, Carbonell, & Falzon, 1993; Branigan, Pickering, Pearson, McLean, & Brown, 2011; Chalnick & Billman, 1988; Hayashi & Miwa, 2009; Pearson, Hu, Branigan, Pickering, & Nass, 2006). An increase in the ratio of lexical alignment was also demonstrated in such scenarios (Branigan et al., 2011).

Branigan et al. (2011) conducted a controlled laboratory-based experiment to investigate how speakers describe the names of objects pictures depending on whether they believe their interlocutor is a human or a computer. The participants and their interlocutors performed the task together. Both look at two pictures; one states the name of the object shown in one of the pictures, and the other selects the picture of the object named by their partner. They repeat this process for various pictures while switching roles alternately (we reuse this approach, see the Task section for details). In their experiments, the participants who were told that their partner was a computer used the name that their partner had used before (i.e., lexical alignment) more often than those who were told that their partner was a human. This decision was interpreted as the participants using the same phrase because they perceived computers as possessing less language knowledge than humans.

Research Questions

In the current study, we developed cognitive models to replicate the results from the experiments in Branigan et al. (2011) and to understand the types of cognitive processes that influence lexical alignment. The adjustments of parameters reveal the relationship between memory retrieval and the top-down process in lexical alignment.

In experiments in Branigan et al. (2011), unexpected names were used to describe a target picture (e.g., "coach" to describe a photo of a bus. Such anomalous and unexpected trials raised the participants' awareness, thereby activating a memory of the trial regardless of the partner's attributes. Unexpected and surprising events are encoded as strong memories (Greve, Cooper, Kaula, Anderson, & Henson, 2017), also known as the von Restorff effect, see Chapter 5 in Ritter, Baxter, and Churchill (2014). Therefore, we tested whether additional activation is necessary for explaining lexical alignment.

We assumed that the top-down process affects the retrieval strategy—that is, the frequency with which the model attempts to retrieve the partner's expressions. If the partners' language proficiency seems low, the speaker should use the word already used to enhance the partners' understanding. Meanwhile, even when the speaker assumes their partner to be an adult human with normal language ability, they still employ lexical alignment. Therefore, we sought to determine if, in such a "neutral" situation where there is a decreased need for careful word selection, the speaker would still display a preference for retrieving the partner's expression rather than using the familiar expression.

The main objective of the study is to identify the causes of the differences between the human–partner and agent partner conditions. First, we tested whether the above two mechanisms—that is, additional activation and retrieval strategies—can produce these differences. If they are possible, based on the differences in the parameter settings, we discuss where the differences in the lexical expression depend-

ing on the partner's attribute derived from.

Human Data

We constructed models that fit the data in Experiments 1, 2, and 4 from Branigan et al. (2011). Below, we briefly describe the task they employed and the results of their experiments.

Task

Their task consisted of two types of trials—a matching and naming trial. While the participants were performing the matching trial, their partner worked on the naming trial and vice versa.

Matching Trial   During the matching trial, the partner performs the naming trial with the corresponding pictures. Initially, two pictures are presented on the screen. After 4000 to 5500 ms, the name of an object is presented, which is then entered by their partner. The participant presses the key corresponding to the picture with the printed name, and a mark appears on the selected picture.

Naming Trial   The naming trial begins with the presentation of two pictures. After 2000 ms, a mark appears above one of the pictures. In Experiments 1 and 4, the participants indicate the name of the object name in the marked picture using a keyboard. This is done orally in Experiment 2. The participants are told that the entered name was sent to their partner, and the partner selects one picture based on the information.

Experimental Conditions   The researcher told half of the participants that their partner was a human and told the other half that their partner was a computer. For both groups, the partner was a computer controlled by the researcher, and it responded in the same manner.

Index   Throughout the experiment, the participants occasionally observed their partner naming an object with an unpopular but correct name, such as "coach" in reference to a bus (this name was defined as a *disfavored* name). Branigan et al. (2011) examined whether participants used an unpopular disfavored name or a common name (it was defined as a *favored* name) when naming the same image. The current study seeks to use their data to simulate the ratio of the trials in which the participants used the disfavored name for an image for which their partner had already used the disfavored name. Hereafter, we refer to this rate as the "use rate."

Data Used for Parameter Fitting

We developed a model that could replicate the use rates in Experiments 1, 2, and 4, as shown in Branigan et al. (2011). There were 16, 32, and 24 participants and 18 disfavored name trials in Experiments 1 and 2, and 16 disfavored name trials in Experiment 4 we used.

In Experiments 1 and 2, there were two trials (one naming and one matching trial) between the trial in which the partner used a disfavored name for an image and the trial in which the participants had to name the same image. The tasks in

One specific name
Knowledge: The name of img1 is name1.
(name1
    isa   img-name
    image img1
    name   name1)

Initial base-level activation
  = ln(1000/(1−0.5)) − 0.5 x ln(10000)
  ≈ 2.996

(a) One specific name

Favored name
Knowledge: The name of img2 is name2a.
(name1
    isa   img-name
    image img2
    name   name2a)

Initial base-level activation
  = ln(800/(1−0.5)) − 0.5 x ln(10000)
  ≈ 2.773

(b) Favored name

Disfavored name
Knowledge: The name of img2 is name2b.
(name1
    isa   img-name
    image img2
    name   name2b)

Initial base-level activation
  = ln(200/(1−0.5)) − 0.5 x ln(10000)
  ≈ 1.386

(c) Disfavored name

Partner's knowledge
Knowledge: The name of img2 for the partner is name2a.
(name1
    isa   img-name
    image img2
    name   name2a
    context partner)

Initial base-level activation
  = ln(n/(1−0.5))
  ≈ 0.693 ~ 3.178
  n is the activation parameter varying from 1.0 to 12.0.
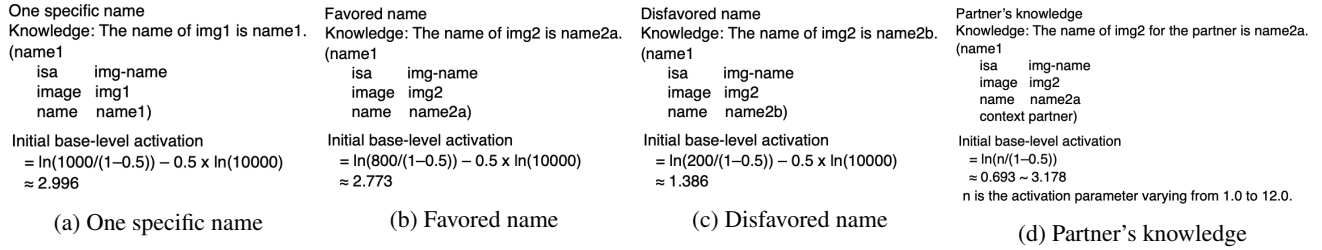
(d) Partner's knowledge

Figure 1: Examples of a chunk of knowledge in the model used in this study. Respectively, they represent when (a) an object has one favored name, an object has both a (b) favored name and (c) disfavored name, and (d) the model obtains a partner's naming.

Experiments 1 and 2 were identical, except that the participants expressed the name of the object with a keyboard for Experiment 1 and orally for Experiment 2. Because there was no significant difference in the use rate between the two experiments, the average use rates for both experiments were used to carry out a fitting process. When participants were instructed that their partner was a computer, they tended to use the disfavored name, with a use rate of 79.5% (Experiment 1: 77%; Experiment 2: 82%). In contrast, when the participants were informed that their partner was a human, the average use rate was 50.5% (Experiment 1: 43%; Experiment 2: 58%).

For Experiment 4, Branigan et al. (2011) manipulated the number of trials between the presentation of the same picture. In one condition (lag 0), an image to which a disfavored name could be given appears in two successive trials; that is, the partner was given a disfavored name during a participant's matching task, and the same picture was presented in the next naming trial. In another condition (lag 8), there were eight trials (four naming and four matching trials) before the second presentation of the image. For cases where the participant believed their partner was a human, there was a significant difference in the use rate between the two lag conditions. The use rate was 54% and 33% for lags 0 and 8, respectively. However, the use rate was not significantly different when the partner was believed to be a computer (lag 0: 76%; lag 8: 82%)[1].

Model

We used ACT–R 7.21 to construct our model (Anderson, 2009). The ACT–R model acquires the current task status through vision and audio modules. The model can add manipulations to the task via the motor and speech modules. Declarative knowledge is stored in and retrieved from the declarative module in form of chunks. Example chunks are presented in Figure 1. Furthermore, a sequence of actions of the model is determined by its production rules. A rule with conditions matching the current status is selected and fired, and actions specified in the rule are conducted.

[1]Refer to Branigan et al. (2011) for more information.

Knowledge in Model

Knowledge of the objects' names was provided in the declarative module in advance. This was based on the participants possessing this knowledge prior to involvement in the experiment. Knowledge was stored in the form of chunks that consisted of slots (an image and name) and their value. Figure 1a provides an example of a chunk ("the name of the img1 is name1"). Each image was given a specific value (e.g., img1) that the model could read when it shifted its attention to the image. Based on this value, the model retrieves the appropriate knowledge and finds the appropriate name of the object in the image. For instance, when the model moves its attention to image1, it can read the value of the image, which is img1. Based on this value, the model can retrieve the knowledge shown in Figure 1a and identify its name, name1. This could be expanded with VisiTor (Tehranchi, Bagherzadeh, & Ritter, 2023) or with SegMan (Amant, Horton, & Ritter, 2007) so that the model might identify objects from visual cues.

For some objects, the knowledge of both favored and disfavored names was provided as different chunks (Figures 1b, 1c). Therefore, two knowledge chunks could be retrieved for one image. Among the two knowledge chunks, the one with the greater activation value was selected. Activation is the value of the base-level activation plus the noise value. In ACT–R, the base-level activation of a chunk i is calculated using the following formula when the optimal learning option is turned on:

$$B_i = \ln\left(n/(1-d)\right) - d \times \ln\left(L\right) \quad (1)$$

where n represents the number of presentations of chunk$_i$, L is the time since the creation of chunk$_i$, and d represents the decay parameter. The activation value increases as the number of presentations increases, which, in turn, decreases as a power function as time passes. We set the decay parameter to the recommended value of 0.5. Noise was generated from a logistic distribution with a mean of 0 and an s set to 0.2.

Humans can retrieve both favored and disfavored names at any time. Therefore, we assigned a high activation value by setting the number of presentations n to 800 for the favored names and 200 for the disfavored names. The difference was due to the favored names being retrieved more easily than the

disfavored names. If the object had only a favored name, then the number of presentations was 1000. The creation time L was set to 10000 to ensure that activation did not decrease as time passed. Examples of this are shown in Figures 1a to 1c. Additionally, the model needed to understand how to pronounce a word to verbalize the object's name. This knowledge was also given to the model with 1000 presentations and a creation time of 10000.

While the model was performing the task, it acquired knowledge about the name that the partner used for the object (partner's knowledge; Figure 1d). Knowledge was represented as a chunk constituted of an image, name, and context slots. To indicate that it was the partner's knowledge, the context slot was assigned the value "partner."

How Our Model Performs Task

When two pictures were presented at the start of each trial, the model encoded each picture and attempted to retrieve each name. First, it retrieved the name with the highest activation value regardless of its context (retrieve-name-right/left). In nearly all cases, the favored name was retrieved. Following this, there were two possible production rules to be selected. In one rule, the retrieved favored name was used as the object's name (name-right/left-img), and in the other, the model attempted to re-retrieve knowledge by constraining its context to a "partner" (retrieve-partner-knowledge-right/left). More specifically, in the latter case, the model attempted to use what its partner had previously stated. The determination of which rule to select was based on its utility value. If the retrieve-partner-knowledge-right/left was selected, and the model successfully retrieves the partner's knowledge, then lexical alignment would occur.

Simulation

Parameters

We varied the values of the two parameters that affect whether the model used the name dictated by its partner.

Utility: Retrieval Strategy Whether the model attempted to use the partner's knowledge depended upon which production rule was fired, "name-right/left-img" or "retrieve-partner-knowledge-right/left." When the conditions of multiple rules match the current state, the rule with the highest utility value is selected. A utility value of 10.0 was given to all rules in advance, and the noise that was generated from a logistic distribution with a mean of 0, and an s of 0.2 was added. Some important rules, such as finding a mark or a name, had a higher utility value (12.0). To change the probability of the rule "retrieve-partner-knowledge-right/left" selected, the utility value for this rule was altered from 9.0 to 11.0 at intervals of 0.2. Additionally, 8.0 and 12.0 were also tested.

Activation: Additional Activation Even if the rule to retrieve the partner's knowledge was fired, the model would still fail to retrieve it if its activation value did not exceed the
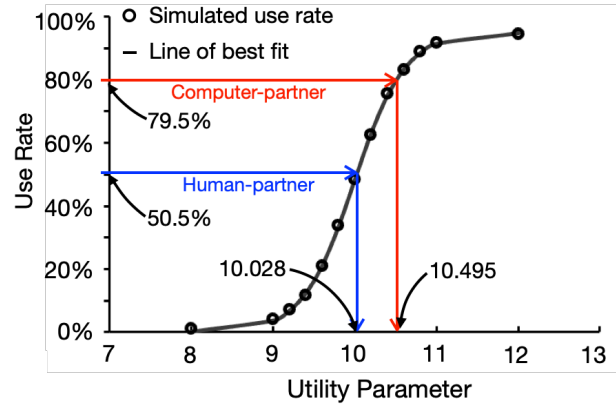

Figure 2: Example result of the first fitting showing the use rates when the value of the activation parameter is set at 4.0. The red line represents how to find the value of the utility parameter that produces the use rate of the computer–partner condition (79.5%) in Branigan et al. (2011), and the blue line represents the use rate in the human–computer condition.

retrieval threshold (set to 0.0 in this model). If the participants intended to use the partner's knowledge in subsequent naming trials, they needed to store the knowledge as important information so that it could be retrieved after a series of trials. Therefore, we manipulated the activation level when the partner's knowledge entered the declarative module. The high initial activation level meant that the model memorized the partner's knowledge.

ACT–R lacks such a mechanism for changing the strength of the initial activation according to its importance. In our model, to change the initial activation level, we manipulated the number of presentations—that is, n (activation parameter) in formula 1—when the chunks of the partner's knowledge enter the declarative module. In the default setting, n was 1.0, and we varied it from 1.0 to 12.0 at intervals of 0.5. For example, if n was 3.0, the initial activation of the partner's knowledge was as high as the activation level when the model encountered this information three times. Due to this increase, the model could retain the activation of the partner's knowledge at a sufficiently high level to ensure successful retrievals in the distant trial.

Fitting to Experiments 1 and 2

First, we searched for parameter values that could reproduce the results of Experiments 1 and 2 from Branigan et al. (2011). For each activation parameter (from 1.0 to 12.0 in 0.5 increments), we ran our model by varying the value of the utility parameter from 9.0 to 11.0 in intervals of 0.2. Moreover, 8.0 and 12.0 were also tested. After performing 100 runs for each combination of parameter values, the use rates formed a sigmoidal function for the utility parameter at each activation level, as shown in Figure 2. A logistic curve with three parameters (Equation 2) was fit to these results us-

(a) Simulated use rate by the human-partner model.

(b) Simulated use rate by the computer-partner model.

Figure 4: Simulated use rate via (a) the human–partner model and (b) the computer–partner model.
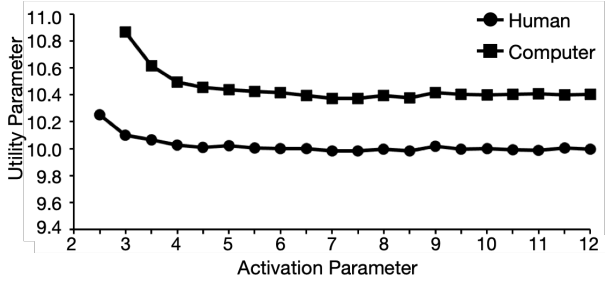

Figure 3: Values of utility parameter producing target use rate. The graph shows utility values that produce the use rate of Branigan et al. (2011) in each activation parameter value. A line with squares is for the human-partner condition and a line with circles is for the computer-partner condition.

ing a simplex function in R^2. The residual was less than 0.005 at all activation levels.

$$use\ rate = \frac{p_1}{1 + exp^{(p_2 + p_3 \times utility)}} \qquad (2)$$

Using the derived logistic formula, we calculated the value of the utility parameter that produced the averaged use rates in Experiments 1 and 2 in Branigan et al. (2011) (human: 50.5%; computer: 79.5%) for each activation parameter level. Figure 3 represents the calculated values of the utility parameter at each activation level. In total, 20 and 19 combinations of the parameter values were discovered for the human and computer conditions, respectively. No utility value could reach the target use rate when the activation parameter was less than 2.5 for the human condition and less than 3.0 for the computer condition. When the activation value was more than 5.5, the utility parameter value converged at approximately 10.0 in the model for the human condition and at approximately 10.4 in the model for the computer condition. The use rate was higher in the computer condition, and therefore, if the value of the activation parameter was the same, then the value of the utility parameter would have been greater in the computer condition model than in the human condition model.

Fitting to Experiment 4

We changed the presentation order of the images to the model to replicate the same lags as those in Experiment 4 from Branigan et al. (2011). Using the combinations of the parameter values found in the first fitting, we ran the model 200 times for each lag 0 and lag 8 task. Figure 4a represents the use rates simulated by the model in each combination of parameter values for the human condition. In Branigan et al. (2011), the use rate was 54% for lag 0, and 33% for lag 8, as represented by the green and pink lines, respectively. Figure 4b shows the results of the models for the computer condition. The use rate reported by Branigan et al. (2011) is shown by the pink line (lag 8: 82%) and the green line (lag 0: 76%), respectively.
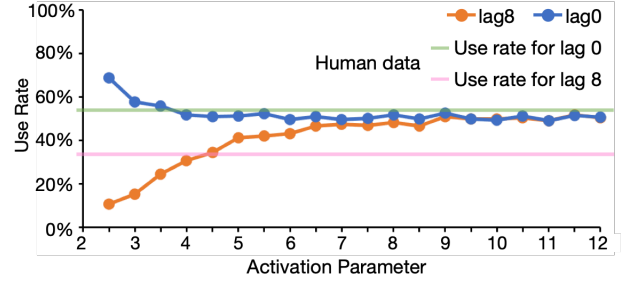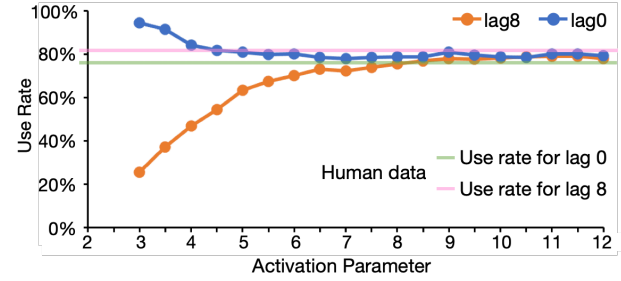
[2]http://aoki2.si.gunma-u.ac.jp/R/simplex.html
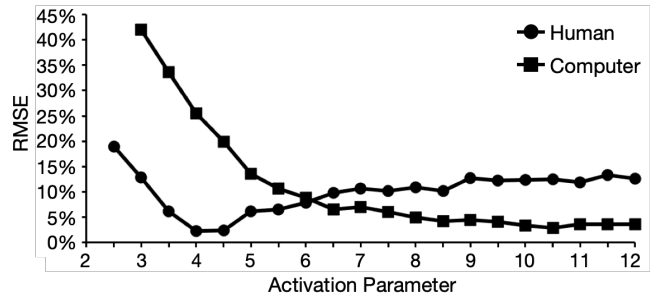
Figure 5: Root mean square error for each combination of parameter values.

We calculated the root mean square error (RMSE) for each combination of parameter values. As suggested in Figure 5, in the human condition, the RMSE was the lowest when the value of the activation parameter was from 4.0 to 4.5, in which the utility value was approximately 10.0. In this range, the use rates were similar to the results reported by Branigan et al. (2011), and the difference between the two lag conditions was highly apparent. In contrast, the RMSE for the computer condition was nearly constant or repressed once the value of the activation parameter exceeded 9.5, in which case the utility value was approximately 10.4. The use rate was approximately 80%, and there was no difference between the two lag conditions.

In summary, the model that was able to explain the results of all lag tasks in the human condition displayed a value of approximately 10.0 for the utility parameter and 4.0 to 4.5

for the activation parameter. For the computer condition, the models with utility parameters of approximately 10.4 and activation parameters higher than 9.5 provided good results. The model for the participants who were instructed that their partner was a computer had the following features compared to the model for those instructed that their partner was a human: (a) The utility for the rule "retrieve-partner-knowledge-right/left" was higher, which led the model to attempt to retrieve the partner's preferred name rather than using the easily retrieved favored name; (b) the model maintained an activation value for partner knowledge that was high enough to be retrieved even after the eight filler trials.

Discussion

This study explored cognitive aspects related to lexical alignment and differences in the cognitive processes of verbal communication between humans and computers. Parameters for the initial activation of the partner's expression and retrieval strategy tendency were adjusted to replicate the results of the human–partner and computer–partner conditions in Experiments 1, 2, and 4 from Branigan et al. (2011). Below, we discuss the findings in the context of our three research questions.

Additional Activation

Even when replicating the results for the human–partner condition, we provided the model with a higher initial activation for the partner's expression. This increased the possibility that the model successfully retrieved those memories when it faced the same picture in the naming trial. Unpredicted expressions attracted attention, required more processing, and raised awareness; therefore, those expressions would have higher availability.

We varied the strength of the initial activation to ensure that the model maintained a high activation for the partner's knowledge. Other strategies, such as rehearsal, can maintain a high activation level. Based on the definition of the activation formula of optimal learning, the activation value was identical when the activation parameter was 9.0 and when the model rehearsed the knowledge nine times. However, we assumed that the participants must have assigned a high activation to the knowledge at the beginning because it was predicted that the information would be used later. Neither previous studies nor our models determined how to acquire a significantly high activation value. Perhaps the participants rehearsed the information very quickly, used memories within the declarative module, or employed some other strategy (cf. Dancy, Ritter, & Berry, 2012; Fum & Stocco, 2004).

Retrieval Strategy

The model for the human–partner condition has nearly the same utility value for the "retrieve-partner-knowledge" and "name-right/left-img" rules. The selection probability for using the familiar name and trying to retrieve the partner's knowledge was at a chance level (50%). This condition is considered the basic situation, and depending on the partner's attribute, the preference between the two rules is adjusted.

Differences between Human and Computer Partners

The model for the computer–partner condition displayed a value for the activation parameters that was more than twice that for the model for the human–partner condition. The computer agent partner was considered to be the one programmed by a developer, and it responded based on an algorithm. Names that did not follow the naming rule for the preceding trials were less anticipated and more surprising than those of the human–partner. Additionally, the participants could have kept those names in their minds as cues to estimate the agent's algorithm.

The utility value for the "retrieve-partner-knowledge" rule was greater in the computer condition model than in the human condition model. Namely, the computer–partner model attempted to retrieve the partner's previous statements rather than use the most popular names that were easily retrieved. In 2011, when Branigan et al. conducted their experiments, language processing was not well developed. Therefore, low faith in the language proficiency of computers would have motivated participants to retrieve their partners' knowledge for immediate successful communication (Cai, Sun, & Zhao, 2021).

The utility value was given at the start of the task depending on the belief about the partner, and it remained constant throughout the task. However, in reality, a change in communication strategy also depends on the partner's actions (Kraut, Lewis, & Swezey, 1982). In Branigan et al. (2011), when the participants believed their partner was a computer, the use rate for the lag 8 condition was higher than that for the lag 0 condition while the difference was not significant. To obtain this result, the utility value for the rule "retrieve-partner-knowledge" must increase during the task. Further work to implement the bottom-up process in our model is needed. As in related studies, researchers showed that users took different processes to evaluate trust in automated systems and human partners, even if both expressed the same action (Madhavan & Wiegmann, 2007). Similarly, the change of the utility would be different depending on whether the partner was a human or a computer; this requires future testing in future research.

Acknowledgment

References

Amalberti, R., Carbonell, N., & Falzon, P. (1993). User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies*, *38*(4), 547–566.

Amant, R. S., Horton, T. E., & Ritter, F. E. (2007). Model-based evaluation of expert cell phone menu interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *14*(1), 1–es.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Baylor, A. L. (2009). Promoting motivation with virtual agents and avatars: role of visual presence and appearance. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3559–3565.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

Cai, Z. G., Sun, Z., & Zhao, N. (2021). Interlocutor modelling in lexical alignment: The role of linguistic competence. *Journal of Memory and Language*, *121*, 104278.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in psychology* (Vol. 9, pp. 287–299). Elsevier.

Dancy, C. L., Ritter, F. E., & Berry, K. (2012). Towards adding a physiological substrate to ACT–R. In *Proceedings of the 21st conference on behavior representation in modeling and simulation* (pp. 78–85).

Fum, D., & Stocco, A. (2004). Memory, emotion, and rationality: An ACT–R interpretation for gambling task results. In *Proceedings of the international conference on cognitive modelling* (pp. 106–111).

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*(3), 378.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition*, *27*(2), 181–218.

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149-165.

Hayashi, Y., & Miwa, K. (2009). Cognitive and emotional characteristics of communication in human-human/human-agent interaction. In *International conference on human-computer interaction* (pp. 267–274).

Kraut, R. E., Lewis, S. H., & Swezey, L. W. (1982). Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, *43*(4), 718.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227.

Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., & Nass, C. I. (2006). Adaptive language behavior in hci: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1177–1180).

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*(4), 633–651.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Ritter, E. F., Baxter, D. G., & Churchill, F. E. (2014). *Foundations for designing user-centered systems: What system designers need to know about people*. Springer.

Tehranchi, F., Bagherzadeh, A., & Ritter, F. E. (2023). A user model to directly compare two unmodified interfaces: a study of including errors and error corrections in a cognitive user model. *AI EDAM*, *37*, e27.

Changes in Time Preference May Simply be Induced by Changes in Time Perception

Arjun Mitra (arjmit@iitk.ac.in)
Department of Cognitive Science, IIT Kanpur
Kanpur, India

Nisheeth Srivastava (nsrivast@iitk.ac.in)
Departments of Cognitive Science and Computer Science, IIT Kanpur
Kanpur, India

Abstract

Present-focused behavior is traditionally studied using models of diminishing utility and varying rates of discounting the future. Recent efforts to curtail time inconsistencies of delay discounting have incorporated subjective time perception into the normative discount function. However, the ramifications of subjective time on inter-temporal choices have not been clearly examined. We simulate time-consistent exponential and time-inconsistent hyperbolic discounting behavior with subjective time to see how the psychological scaling of objective clock time affects people's choice of the delayed reward. Our results suggest that time contraction and dilation respectively increase and decrease the probability of choosing the later outcome. We also find that these time perception-based preference shifts are similar in effect size to preference shifts typically explained by changes in discount rates earlier in the literature. Our results suggest that a psychological time-perception account can be used to explain observed present-focused behaviors instead of relying on traditional discount-rate explanations.

Keywords: time dilation; time contraction; delay discounting; inter-temporal choices

Introduction

Inter-temporal choices encompass decisions whose consequences play out over time. These decisions are ubiquitous and are often studied as two alternative choices - one rewarding choice that one can get *now* vs. another better reward that manifests over the *future*. A willingness to forgo the sooner reward in consideration for the more significant, later reward is often associated with higher patience or self-control. Such willingness has been empirically tested using tasks like the 'marshmallow task' in kids (Mischel, 2014) or using pairwise monetary comparison tasks spanning different periods of time in adults (Andersen, Harrison, Lau, & Rutström, 2008). This ability to delay gratification is often considered a predictor of higher scholastic abilities, better coping with stress and frustration (Mischel, Shoda, & Rodriguez, 1989), and better self-regulatory behaviors (Michaelson & Munakata, 2020).

Inter-temporal choices involve trade-offs between the costs and benefits of rewards available right now and sometime in the future. Samuelson's 'Discounted Utility (DU)' model first formulated a decision maker's inter-temporal preference using a utility function $U(T)$ which signifies the value the observer assigns to a reward achievable in a distant time. This is mathematically represented as

$$U(T) = \sum_{t=0}^{T-t} f(n) \times U(t) \qquad (1)$$

where $f(n)$ is the discount function, i.e., the decision maker's relative weight assigned to the future reward at time T. According to the DU model, this discount function is exponential $f(n) = e^{-kt}$, and the utility of any future goal u_t at time t is given by

$$u_t = r_t \times e^{-kt} \qquad (2)$$

where r_t is the actual reward at time t, and k is the decision maker's discount factor.

The DU model presupposes that people discount the future in a time-consistent manner, i.e., the discount rate k is fixed over time. However, empirical evidence suggests that people usually discount the future more when the alternative is presented *now* compared to when it is presented after some delay, making future discounting time-inconsistent (Thaler, 1981). To account for this, some researchers above proposed that the discount function $f(n)$ be hyperbolic in nature (Mazur, 2013) such that $f(n) = 1/(1+kt)$. Thus, in hyperbolic discounting, the utility of a future reward u_t is given by

$$u_t = \frac{r_t}{(1+kt)} \qquad (3)$$

where the discount factor k varies with time, yielding more discounting in smaller delays than larger ones.

Why do people often choose the smaller, sooner reward instead of the larger, later one? As formalized by the DU model, if one is not motivated to wait for later, i.e., has a high discount factor, they would perceive the utility of later reward to be smaller and consequently they would opt for the sooner reward. On the other hand, as the delay to reward delivery increases, the utility associated with waiting also decreases. For example, a kid willing to wait seven minutes for two pretzels instead of one might not want to wait fifteen minutes. Various factors like anticipation of a promising event (Loewenstein, 1987), dread of a painful outcome (Berns et al., 2006), cue-induced reward overestimation (Jędras, Jones, & Field, 2014), visceral influences (Loewenstein, 1996), emotional arousal (Lempert, Johnson, & Phelps, 2016), environmental reliability (Kidd, Palmeri, & Aslin, 2013), negative income shocks (Haushofer, Schunk, & Fehr, 2013) has been shown to affect future choices by decreasing the utility of delayed rewards or increasing their discount rates. However, an often overlooked dimension in explaining delay discounting phenomena is the delay itself.

In any discounting model, delay is typically measured in terms of clock time. However, recent explorations into how people perceive time delays reveal interesting insights. McGuire and Kable (2012) have empirically demonstrated that when the delay in inter-temporal choices seems to be increasing over time (like waiting for a phone call) compared to being diminishing over time (like waiting for a bad movie to end), people show preference reversals - they often prefer the delayed rewards initially and then forgo it later. This insight highlights how our perception of inter-temporal delays can affect our choices and can help us identify why people often forego more significant, later rewards. On a similar note, Takahashi (2016) show that if this perceived delay is assumed to be non-linear (logarithmic as in psychophysical experiments) instead of an objective linear time, the exponential discounting function often takes the form of a hyperbolic one. In support of this, researchers have shown empirically that perceived time is indeed non-linear and concave in nature, and that people demonstrate a constant discount rate when subjective time perception is taken into account (Zauberman, Kim, Malkoc, & Bettman, 2009).

If people perceive time non-linearly, how would this psychological scaling of time affect their inter-temporal choices compared to objective time? Intrinsic utility of any reward or the discount rate of an individual is often an immeasurable quantity. Can a mental account of time give a better explanation for delay discounting behavior?

In this article, we incorporate subjective perceived time in delay discounting models to understand how time dilation (when perceived time is > objective time) or contraction (when perceived time is < objective time) can affect intertemporal choices. To be precise, we incorporate different values of wait-time (modeled as subjective time lesser or greater than objective time) in exponential and hyperbolic discounting models to see how preference for later rewards change. Thus, our goal in this paper is to quantify how these deviations from objective time can change the probability of choosing later rewards and to check if these time-warped preference shifts can account for changes in discount rates when objective time is considered. Our methods and their corresponding results are described below.

Intertemporal choice modeling with subjective time

The DU model suggests that people discount future outcomes exponentially based on their discount rates and the delay associated with the outcome. As shown in Eqn 2, as the delay increases, the utility of the future reward decreases. What happens if we replace the objective delay with subjective perceived time? Following Takahashi (2005)'s direction, if we assume mental time to be represented in a non-linear manner following Weber-Fechner's law, the relationship between subjective time t_s and objective time t_o should look like this:

$$t_s = \alpha \times ln(1 + \beta \times t_o) \qquad (4)$$

where α and β are free parameters independent of t_s and t_o. Substituting this subjective time for objective time in Eqn 2, we get

$$u_t = r_t \times exp(-k(\alpha \times ln(1 + \beta \times t_o))) \qquad (5)$$

Rearranging the Eqn 5,

$$u_t = r_t \times exp(ln(1 + \beta \times t_o)^{-k\alpha})$$
$$= \frac{r_t}{(1 + \beta \times t_o)^{k\alpha}}$$
$$= \frac{r_t}{(1 + \beta \times t_o)^s}$$

where, $s = k\alpha$. Thus, Eqn 5, which includes an exponential discount function with logarithmic perceived time, turns into a general hyperbolic function, and if we consider $s = 1$, it turns into a simple hyperbolic function similar to Eqn 3.

The dynamic inconsistency often found in the discounting literature is mitigated by considering mental time representation. It is known that substance abusers often discount delayed rewards more than non-drug dependent subjects, and a hyperbolic discount model often fits the data better than an exponent, time-consistent one (Bickel & Marsch, 2001). In that case, representing time in a non-linear, logarithmic fashion instead of a linear one, as shown above, removes the inconsistency (Takahashi, 2005).

If people represent mental time non-linearly, how do these deviations from objective time affect the probability of choosing the later reward? Assume one has to wait for a year for some reward, and the probability of waiting is p. If mentally that one year feels like a year and a half (time dilation) or six months to them (time contraction), our model formulates how their probability of choosing the later reward p' would change compared to p. Thus, we find how deviations in time $\delta(t)$ modulate deviations in choices $\delta(p)$ using exponential and simple hyperbolic functions.

Exponential discounting

Imagine an agent is faced with two choices - a sooner, smaller reward r_0 and a later, larger reward a_t separated by objective, calendar time t_o. The probability of them choosing the later reward is $p(later)$. Since discount factor k is unknown, we can estimate k given the actual value of the later reward, time to fruition, and the utility associated with it $u(later)$. This $u(later)$ is calculated using a softmax function, which can be represented as

$$p(later) = \frac{exp(u(later))}{exp(u(later)) + exp(u(sooner))} \qquad (6)$$

where $u(sooner)$ is the utility associated with the sooner reward, which is assumed to be equal to r_0. Rearranging Eqn 6, we get

$$exp(u(later)) = p(later) \times exp(u(later)) + p(later) \times exp(u(sooner))$$
$$exp(u(later)) = \frac{p(later) \times exp(u(sooner))}{1 - p(later)}$$

Thus, if we know $p(later)$, we can derive the utility of the later reward $u(later)$ at time t_o using

$$u(later) = ln(\frac{p(later) \times exp(u(sooner))}{1 - p(later)}) \quad (7)$$

The exponential discount function, given by Eqn 2, can be rearranged in our context to give

$$u(later) = a_t \times exp(-k \times t_o)$$

$$exp(-k \times t_o) = \frac{u(later)}{a_t}$$

$$k \times t_o = ln(\frac{a_t}{u(later)})$$

Given $u(later)$ obtained from Eqn 7, and $p(later)$, we can derive our agent's discount factor k using objective time t_o and the actual later reward value a_t using the following formula

$$k = \frac{1}{t_o} \times ln(\frac{a_t}{u(later)}) \quad (8)$$

Now, if our agent mentally represents objective time t_o subjectively as t_s, we can find the updated utility of the later reward $u(later)'$ given t_s using k from Eqn 8

$$u(later)' = a_t \times exp(-k \times t_s) \quad (9)$$

And using this $u(later)'$, we can find the new probability of choosing the later reward $p(later)'$ using the softmax function

$$p(later)' = \frac{exp(u(later)')}{exp(u(later)') + exp(u(sooner))} \quad (10)$$

Finally, we can quantify how deviation in time $\delta(t)$ can perturb the probability of choosing later outcome $\delta(p)$ such that

$$\delta(t) = t_s - t_o \quad (11)$$

$$\delta(p) = p(later)' - p(later) \quad (12)$$

For our model, our agent can choose from a sooner reward $r_0 = 100$ available at time $t_o = 0$ or wait a year $t_o = 365$ for a reward of $a_t = 150$. Given different probabilities of choosing the later reward $p(later)$ ranging from 0.1 to 0.9, we calculate $u(later)$ using Eqn 7 and k using Eqn 8. Assuming that subjectively waiting for a year could feel like waiting for six months i.e., $t_s = 180$ days (time contraction by six months) or waiting for a year and a half i.e., $t_s = 545$ days (time dilation by six months), we calculate the perceived utility of later reward $u(later)'$ using Eqn 9 and the updated probability of choosing the later reward $p(later)'$ using Eqn 10. From this, we find $\delta(t)$ and $\delta(p)$ using Eqns 11 and 12 to check how $\delta(p)$ changes as a function of $\delta(t)$.

We find that $\delta(p)$ changes in a sigmoidal manner as a function of $\delta(t)$. In Fig 1, the dotted line corresponding to $\delta(t) = 0$ signifies subjective time being equal to the objective time, the negative x-axes signify the perceived shortening of time
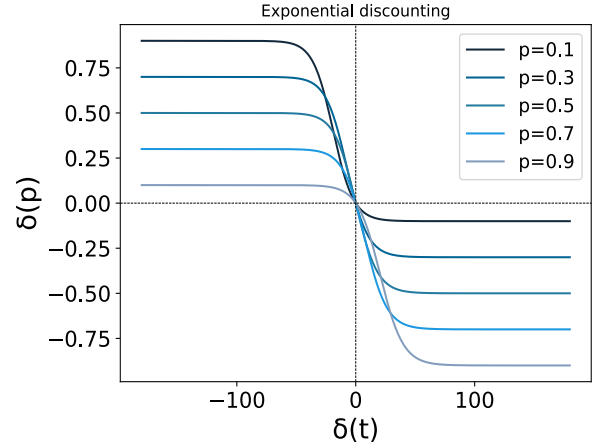

Figure 1: A figure depicting how deviations in time $\delta(t)$ perturb the probability of choosing a later reward $\delta(p)$ when the future reward is discounted exponentially. We find that as time dilates (i.e., $\delta(t) > 0$ such that subjective time > objective time), the probability of choosing the future outcome decreases (i.e., $\delta(p) < 0$ such that $p(later)$ at subjective time $< p(later)$ at the objective time). For this simulation, we assumed actual later reward $= 1.5 \times$ sooner reward.

(i.e., time contraction), and the positive x-axes signify the perceived lengthening of time (i.e., time dilation). As time contracts ($\delta(t) < 0$) and time dilates ($\delta(t) > 0$), we see a rise ($\delta(p) > 0$) and fall ($\delta(p) < 0$) in the probability of choosing later rewards respectively for all values of prior probability p ranging from 0.1 to 0.9. As time dilates, this fall in probability is maximum when the prior probability is high ($p = 0.9$) and minimum when it is low ($p = 0.1$), as shown in the fourth quadrant of Fig 1. Thus, our agent's preference for later rewards significantly falls when mental time dilates, corresponding to objective time. This fall is proportional to their prior probability of choosing the later reward - as their prior probability grows higher (p goes from 0.1 to 0.9), their shift in preference also grows steeper. This seems intuitively logical - if one prefers to delay gratification significantly but their wait time seems to be extending in their mind, the subjective utility of that later outcome decreases, leading to a drop in their probability of choosing that reward. Thus, instead of waiting, they may reverse their preference at some point in time and choose the smaller reward.

To check the robustness of our model, we varied the value of the later reward and found that the same results were reproduced as shown in Fig 2. Whether we make the value of the later reward smaller than our original model (Fig 2(a)) or larger (Fig 2(b)), we find that as time dilates ($\delta(t) > 0$), the agent's preference for later reward decreases. However, the nature of this descent is slower when the later reward is 1.1 times that of the sooner reward, as shown in Fig 2(a). When the prior probability is low ($p = 0.1$) and as time dilates ($\delta(t) > 0$), the $\delta(p)$ decreases marginally below 0 and
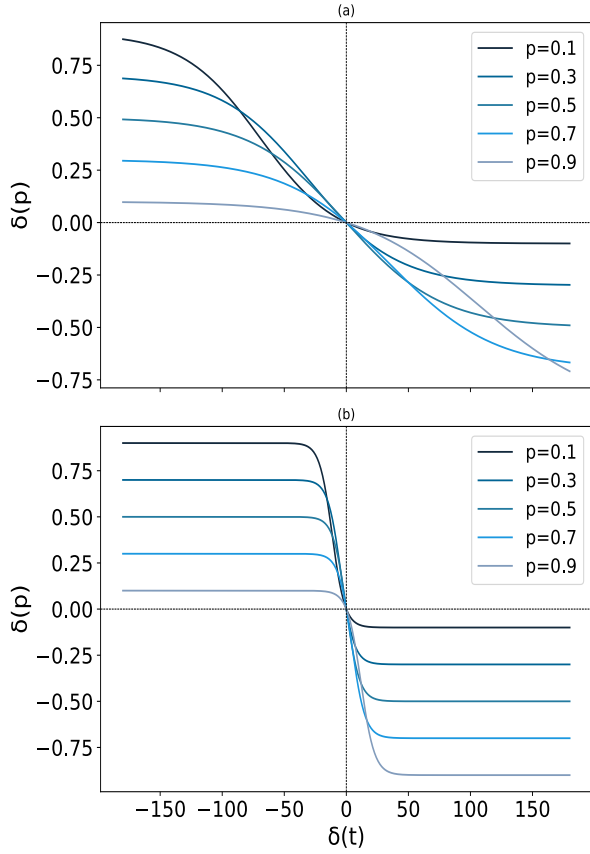
Figure 2: This figure depicts how varying the values of later rewards made the same predictions as we had previously found. The plot (a) and (b) shows simulation results for conditions where the actual value of the later reward is taken to be smaller (1.1 × sooner reward) and larger (2 × sooner reward) than the one used in the main simulation (1.5 × sooner reward).

quickly asymptotes for all values of later reward.

On the other hand, the decrease in $\delta(p)$ is significantly more when the prior probability of choosing the later reward is high ($p = 0.9$) compared to when it is low ($p = 0.1$). If we compare all values of later reward as seen in Fig 1 and 2, we find that the point in time where $\delta(p)$ asymptotes gets smaller as the value of later reward increases ($\delta(t) > 180$ in Fig 2(a), $\delta(t) \approx 50$ in Fig 1, and $\delta(t) \approx 25$ in Fig 2(b) for $p = 0.9$). This trend continues for other values of delayed rewards that are more than twice the size of the sooner reward.

Hyperbolic discounting

We followed the same protocol as above, but instead of using an exponential function, we used a simple hyperbolic function to define the utility of the later reward given by

$$u(later) = \frac{a_t}{1 + k \times t_o} \tag{13}$$

where k is the discount factor, a_t is the actual reward manifesting at objective time t_o. Rearranging this eqn, we get the discount factor where

$$k = \frac{a_t - u_t}{u_t \times t_o} \tag{14}$$

In this simulation, inter-temporal choices are also defined as a sooner reward $r_0 = 100$ available at $t_o = 0$ and a delayed reward $a_t = 150$ redeemable at $t_o = 365$. Given different values of $p(later)$ ranging from 0.1 to 0.9, we estimate the $u(later)$ using Eqns 7. Then using this $u(later)$, we estimate k using Eqn 14. Using this k and plugging subjective time $t_s = t_o + \delta(t)$ in Eqn 13, we estimate $u(later)'$ and finally $p(choice)'$ using Eqn 10. Lastly, we find the deviations in time and probability $\delta(t)$ and $\delta(p)$ using Eqns 11 and 12. Like the exponential case, we find $\delta(p)$ to be changing sigmoidally as a function of $\delta(t)$ as seen in Figure 3.
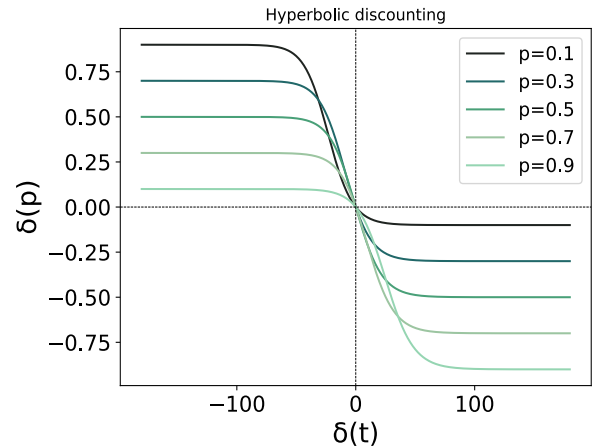
Figure 3: This figure shows how deviations in time $\delta(t)$ modulate the probability of choosing a later reward $\delta(p)$ when the future reward was discounted hyperbolically. We find that as time dilates (i.e., $\delta(t) > 0$ such that subjective time > objective time), the probability of choosing the future outcome falls ($\delta(p) < 0$) for all values of prior probability p. For this simulation, we also assumed actual later reward = 1.5 × sooner reward.

Similar to the exponential discount scenario, we find that as time dilates ($\delta(t) > 0$) such that subjective time is perceived to be longer than objective time, the probability of choosing the later reward decreases ($\delta(p) < 0$). This descent was highest when the prior probability of choosing the later reward was high and vice versa. Similarly, as time contracts ($\delta(t) < 0$) such that subjective time is smaller than clock time, the choice of the delayed reward increases ($\delta(p) > 0$).

We also performed robustness checks of our results by varying the size of the later reward. For both lower and higher values of later reward than our original model, we found that as time dilates ($\delta(t) > 0$), the probability of choosing later rewards also decreases ($\delta(p) < 0$). For high values of prior

If that is true, delay discounting behavior can be explained in a quantifiable mental-time model compared to an immeasurable discount rate.

Exponential discounting

In the above section, we estimated our agent's discount factor k for each level of prior probability of choosing later rewards. We used that to assess how this probability changed in the face of deviations from objective time. Now, if we disregard those time deviations and consider time to be objective and constant, the changes in preference would, *ceteris paribus*, appear to correspond to changes in the delay discounting parameter.

Figure 4: This figure depicts the robustness check performed for the change in $\delta(p)$ as a function of $\delta(t)$ for the hyperbolic discount function. The plot (a) shows simulation results for conditions where the later reward is smaller ($1.1 \times$ sooner reward) than the main simulation ($1.5 \times$ sooner reward). The plot (b) shows simulation results when the later rewards were larger ($2 \times$ sooner reward) than the main simulation.

Figure 5: This figure shows how discount rates change $\delta(k)$ as a result of observed shifts in preference $\delta(p)$ if the psychological scaling of objective time is disregarded in an exponential discounting model. For all prior proba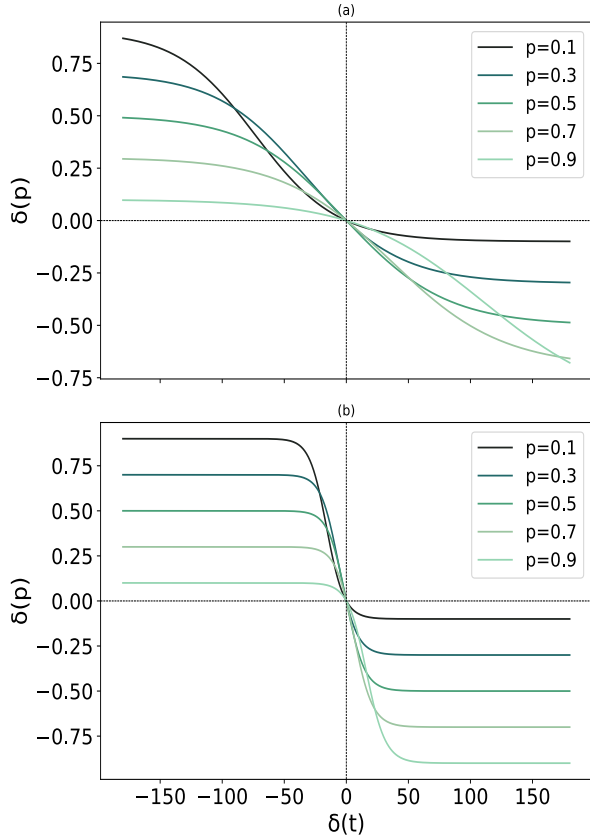bility values, as the preference for the delayed rewards decreases, the discounting increases when only objective time is considered.

probability $p = 0.9$, the fall in probability ($\delta(p)$) is much more gradual when the later reward is 1.1 times the sooner reward compared to when it is twice as big as the sooner reward. Again we find the point in time where $\delta(p)$ asymptotes get smaller as the value of later reward increases, as can be seen in Fig 4(a), 3, and 4(b).

Time dilation may explain delay discounting

In the previous section, we varied the time parameter in the discounting models to see how the probability of choosing later rewards changed while keeping the discount rate constant. We find that our agent's preference for delayed reward decreases across exponential and hyperbolic discounting formulations as the perception of time lengthens compared to objective clock time. This leads us to ask whether these time-warp-induced preference changes can explain changes in present-focused behavior. If we assume time to be objective and non-variable, do these shifts in the probability of choosing later rewards translate to changes in discount rate?

To test this possibility, we use the observed changes in the probability of choosing later rewards $\delta(p)$ as a result of time deviations (as shown in Fig 1) to calculate $p(later)'$ using Eqn 12. Using this $p(later)'$, we calculate the utility associated with the later reward $u(later)'$ using a softmax function as shown in Eqn 7. Assuming perceived time to be similar to the objective time ($t_s = t_o$), we calculate the discount factor k' (using an exponential discounting model) for each observed change in utility using

$$k' = \frac{1}{t} \times ln(\frac{a_t}{u(later)'})$$

where time $t = t_s = t_o$. We quantify the changes in discount rates by

$$\delta(k) = k' - k \tag{15}$$

where k is calculated using objective time t_o and actual later reward a_t using Eqn 8 for all values of prior probability p. Lastly, we plot how discount rates change $\delta(k)$ as a function of our observed changes in preference of delayed reward $\delta(p)$

if subjective scaling of time is disregarded and clock time is considered.

As shown in Fig 5, we find that as the preference for later reward decreases (signified by $\delta(p) < 0$), the discount rate increases (signified by $\delta(k) > 0$). Since the preference drop increases as the p goes from 0.1 to 0.9, the increase in discount rates is highest for $p = 0.9$ and lowest for $p = 0.1$. Overall, this aligns well with our intuition that when time is treated as objective in modeling intertemporal choice, underlying subjective changes in time perception may well be measured as shifts in discount rates.

Hyperbolic discounting

To check these results' robustness, we performed a similar modeling approach of mapping preference shifts due to temporal deviations to discount rates with hyperbolic discounting.
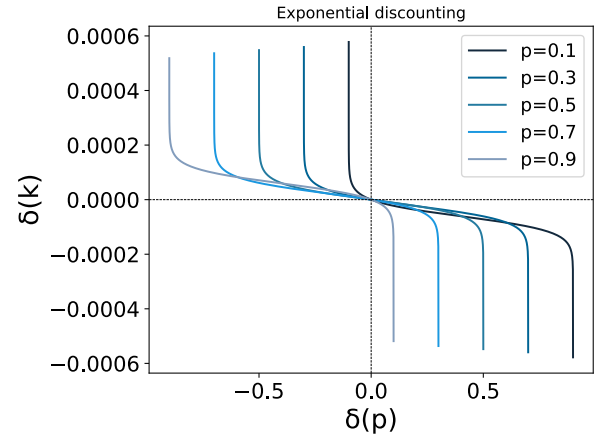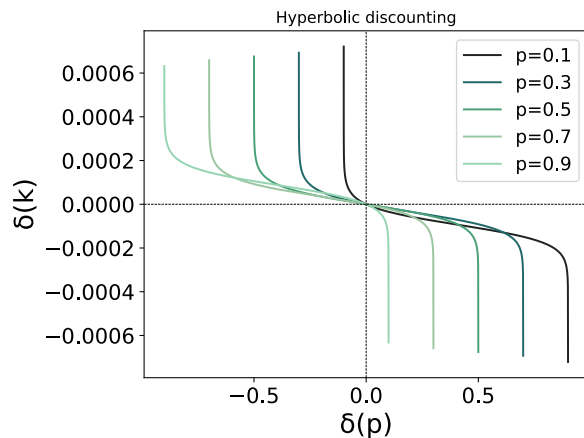

Figure 6: This figure shows how discount rates change $\delta(k)$ as a result of observed shifts in preference $\delta(p)$ if objective time is considered in a hyperbolic discounting model. For all values of prior probability p, as the preference of the delayed rewards decreases, the discounting increases when objective time is considered.

Our protocol was the same as above, except that to find k', we used

$$k' = \frac{a_t - u(later)'}{u(later)' \times t}$$

where time $t = t_s = t_o$. We quantify the changes in discount rates $\delta(k)$ using Eqn 15. We found our results to be exactly similar to the exponential case as shown in Fig 6. These observations suggest that observed shifts in preference often attributed to differential rates of discounting in choice paradigms may well be actually caused by shifts in temporal perception.

Discussion

In line with the new-found interest in understanding discounting behavior in terms of psychologically perceived time, we demonstrated using simulations how changes in time preference conventionally attributed to changes in discount rates may actually be produced by changes in time perception.

Across both models of time-consistent exponential and time-inconsistent hyperbolic discounting, we find a sigmoidal change in preference for delayed rewards as a function of time deviations - when subjective time contracts, the probability of choosing the later reward increases and when subjective time dilates, the probability decreases compared to the prior probability. This seems intuitive - if one perceives a month to be a week, then waiting for a month seems easier and highly likely. However, if waiting for the same month seems like a year, then choosing to wait seems highly unlikely.

We also found that these shifts in probability correspond to changes in discount rates when time is assumed to be objective and constant. For both exponential and hyperbolic discounting, the decrease in the likelihood of choosing a later reward (corresponding to an increase in perceived time) translates to an increase in discount rates when time is considered non-variable. This demonstrates how a mental time narrative can explain discount rate accounts of time preference shifts.

Exponential discounting functions assume discount rates to be constant over time and cannot account for preference reversals (Thaler, 1981; Kirby & Herrnstein, 1995). By incorporating subjective time into exponential models, our in-silico demonstrations suggest a simple explanation: as the perceived time phenomenologically lengthens in comparison to clock time, the favorability of delaying gratification decreases and eventually drops to null - thus explaining preference reversals. In other words, even though one might prefer a long-term reward initially given a description of the anticipated delay, they can switch to a short-term plan if the experience of the delay feels longer, as the delayed outcome might not look lucrative enough on the stretched out subjective timeline.

Understanding the interplay of uncertainty in one's environment, how time is perceived, and how it leads to preference is essential for understanding why people discount the future. Often as ambiguity increases, people's phenomenological experiences intensify, and time seems to linger on (Maglio & Kwok, 2016). Manipulations of perceived control of one's actions and their outcomes distort people's duration judgments of negative images (Mereu & Lleras, 2013), and these time distortions can be subsequently restored by experiences of higher control (Buetti et al., 2020). Thus, if internal time is malleable to our lived experiences, studying time preferences using this prism may yield an enhanced understanding of present-focused behavior in light of this psychological scaling of time. Our model implies that latent traits like impatience or lack of self-control need not be evoked to explain such discounting behavior. Psychological scaling of clock time offers similar explanations and paints delay discounting as an ecologically rational strategy - there is no point in waiting for tomorrow if tomorrow seems like forever.

References

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583-618. doi: https://doi.org/10.1111/j.1468-0262.2008.00848.x

Berns, G. S., Chappelow, J., Cekic, M., Zink, C. F., Pagnoni, G., & Martin-Skurski, M. E. (2006). Neurobiological substrates of dread. *Science*, *312*(5774), 754-758. doi: 10.1126/science.1123721

Bickel, W. K., & Marsch, L. A. (2001). Toward a behavioral economic understanding of drug dependence: delay discounting processes. *Addiction*, *96*(1), 73-86. doi: https://doi.org/10.1046/j.1360-0443.2001.961736.x

Buetti, S., Xue, F., Liu, Q., Hur, J., Ng, G. J. P., & Heller, W. (2020). Perceived control in the lab and in daily life impact emotion-induced temporal distortions. *Timing & Time Perception*, *9*(1), 88–122. doi: https://doi.org/10.1163/22134468-bja10018

Haushofer, J., Schunk, D., & Fehr, E. (2013). Negative income shocks increase discount rates.

Jędras, P., Jones, A., & Field, M. (2014). The role of anticipation in drug addiction and reward. *Neuroscience and Neuroeconomics*, *3*, 1-10.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, *126*(1), 109–114. doi: https://doi.org/10.1016/j.cognition.2012.08.004

Kirby, K. N., & Herrnstein, R. J. (1995). Preference reversals due to myopic discounting of delayed reward. *Psychological Science*, *6*(2), 83–89.

Lempert, K. M., Johnson, E., & Phelps, E. A. (2016). Emotional arousal predicts intertemporal choice. *Emotion*, *16*(5), 647. doi: https://doi.org/10.1037/emo0000168

Loewenstein, G. (1987, 09). Anticipation and the Valuation of Delayed Consumption. *The Economic Journal*, *97*(387), 666-684. doi: 10.2307/2232929

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *65*(3), 272-292. doi: https://doi.org/10.1006/obhd.1996.0028

Maglio, S. J., & Kwok, C. Y. (2016). Anticipated ambiguity prolongs the present: Evidence of a return trip effect. *Journal of Experimental Psychology: General*, *145*(11), 1415. doi: https://doi.org/10.1037/xge0000228

Mazur, J. E. (2013). An adjusting procedure for studying delayed reinforcement. In *The effect of delay and of intervening events on reinforcement value* (pp. 55–73). Psychology Press.

McGuire, J. T., & Kable, J. W. (2012). Decision makers calibrate behavioral persistence on the basis of time-interval experience. *Cognition*, *124*(2), 216–226. doi: https://doi.org/10.1016/j.cognition.2012.03.008

Mereu, S., & Lleras, A. (2013). Feelings of control restore distorted time perception of emotionally charged events. *Consciousness and Cognition*, *22*(1), 306–314. doi: https://doi.org/10.1016/j.concog.2012.08.004

Michaelson, L. E., & Munakata, Y. (2020). Same data set, different conclusions: Preschool delay of gratification predicts later behavioral outcomes in a preregistered study. *Psychological Science*, *31*(2), 193-201. (PMID: 31961773) doi: 10.1177/0956797619896270

Mischel, W. (2014). *The marshmallow test: Understanding self-control and how to master it*. Random House.

Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science*, *244*(4907), 933-938. doi: 10.1126/science.2658056

Takahashi, T. (2005). Loss of self-control in intertemporal choice may be attributable to logarithmic time-perception. *Medical Hypotheses*, *65*(4), 691-693. doi: https://doi.org/10.1016/j.mehy.2005.04.040

Takahashi, T. (2016). *Loss of self-control in intertemporal choice may be attributable to logarithmic time-perception*. Springer.

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, *8*(3), 201-207. doi: https://doi.org/10.1016/0165-1765(81)90067-7

Zauberman, G., Kim, B. K., Malkoc, S. A., & Bettman, J. R. (2009). Discounting time and time discounting: Subjective time perception and intertemporal preferences. *Journal of Marketing Research*, *46*(4), 543-556. doi: 10.1509/jmkr.46.4.543

Trait Inference on Cognitive Model of Curiosity:
Relationship between Perceived Intelligence and Levels of Processing

Kazuma Nagashima (nagashima.kazuma.16@shizuoka.ac.jp),
Junya Morita (j-morita@inf.shizuoka.ac.jp) ,
Department of Informatics, Graduate School of Science and Technology, Shizuoka University,
3-5-1 Johoku, Chuo-ku, Hamamatsu-shi, Shizuoka-ken, 432-8011 Japan

Abstract

Cognitive models are used as simulators that derive external behavior from assumed internal states. As a tool for linking external behavior with internal causes, cognitive models can be used to examine human trait inference on others. While fundamental attribution errors are identified in social psychology, the specific factors remain unclear. By employing detailed cognitive models to specify internal states, it is possible to deepen our understanding of human inference on internal processes. In this study, we utilized the ACT-R cognitive architecture to construct such internal states and externalized behaviors. We also focused on "curiosity" as an individual trait emphasized in real society to evaluate individuals. We developed a visualizer for the behavior of multiple models of curiosity and conducted subjective evaluations with participants recruited from a Japanese crowdsourcing site. As a result, we observed differences in inferred traits among models, although the specific patterns were not consistently aligned with the model assumptions. Additional analysis revealed that participants' inferences were more influenced by observable behavior patterns rather than internal processes, indicating a deficit in human attribution as suggested by the tradition of social psychology.

Keywords: trait inference; subjective human evaluation; cognitive modeling; ACT-R

Introduction

In the community, many researchers have so far developed cognitive models as simulators that derive external behaviors from assumptions about human internal states. Depending on the representations of internal states and external behaviors, a variety of cognitive models at different granularities can be constructed. Each model aims to replicate various errors exhibited by humans while also representing the internal processes occurring in the human brain. Such models serve as representations of theories in cognitive science and are also utilized to predict and interpret human behavior in various contexts such as education and industry.

Based on the above perspective of the history of cognitive modeling, this paper proposes to utilize them as tools for studying human trait inference for others. Humans tend to interpret others' behaviors by attributing internal states such as personality (Winter & Uleman, 1984). Previous research in social psychology has pointed out that such trait inference inherently involves fundamental attribution error (biases) (Gilbert & Malone, 1995). However, research on these biases traditionally relied on surveys and experiments, lacking clear manipulation of internal states that underlie behavior. By applying cognitive models to this area of study, it

becomes possible to detail the factors that lead to errors from the perspective of internal mechanisms causing human behaviors. The outcomes of such studies will be applied to the development of artifacts that interact smoothly with humans, as well as to applications such as the assessment of others in education and work environments.

Among various human internal states, motivation is often attributed as a cause of behavior and holds significant influence. In Weiner (1985)'s attribution theory, motivation is classified as an internal and controllable factor. Attribution to this factor significantly affects the occurrence of challenging behavior in the future. Thus, examining the process of attributing motivation to behavior has both theoretical and practical significance. This study specifically investigates the inference of attributes based on individuals' curiosity, a form of intrinsic motivation. Curiosity has been classified as a form of intrinsic motivation and has been increasingly addressed by numerous computational models (Aubret, Matignon, & Hassas, 2019). Therefore, this study utilizes behavior generated by models of curiosity as experimental stimuli to explore factors influencing its attribution.

Related Works

This study examines whether it is possible to infer the traits of agents with internal models of curiosity by participants. In the aim, we introduce research related to (1) psychological or computational studies on trait inference, (2) studies focusing on internal models of curiosity, and (3) research on evaluating the characteristics of artificial agents with models.

Human Trait Inference and Attribution

The inference of others' attributes discussed in the previous section is related to Theory of Mind (Premack & Woodruff, 1978), the ability of humans to understand others. In connection with this process, people have a strong tendency to find intentionality and animacy in artifacts. The classical study by Heider and Simmel (1944) demonstrated that humans perceive intentions from objects represented by simple geometric shapes.

Such an inference on the inside process frequently connects an individual's traits such as characteristics and thinking style. Regarding this issue, research in social psychology suggests the inaccuracy of human traits inference (Stangor & Walinga, 2014). While humans spontaneously infer traits

from behavior, numerous attribution errors, such as overestimating others, have been reported (Winter & Uleman, 1984). Particularly, humans may perceive intent even in random behavior (Fyfe, Williams, Mason, & Pickup, 2008). However, similar to the classical study by Heider and Simmel (1944), the experimental stimuli used in these experiments are constructed manually lacking details of internal processes.

Based on the background discussed so far, the current study investigates human trait inference using cognitive models. We assume the internal processes of the cognitive model as traits (thinking style) and use the behaviors generated from these traits as experimental stimuli. This manipulation aims to describe human trait inference in terms of computational algorithms.

As mentioned in the introduction, this study focuses on curiosity as the trait for participants to estimate. Curiosity plays a significant role in promoting individual activities across various fields such as education and entertainment.

Because of its importance, many psychological studies have been conducted to clarify the concept of curiosity. For example, Malone (1981) distinguished perceptual curiosity and cognitive curiosity in his theory of intrinsic motivation. Kashdan et al. (2018) also classified multiple aspects of curiosity, such as exploration and absorption, and invented the questionnaire assessing individual traits toward such different aspects of curiosity.

Such psychological theories have recently been elaborated by computational models of human curiosity. Dominant theories are based on principles of prediction errors (Friston, 2010; Schmidhuber, 2010), leading to research on deep learning agents representing curiosity (Aubret et al., 2019). These agents offer a solution to the exploration and exploitation dilemma in agent learning, enhancing performance in specific environments (e.g., games).

One limitation of deep learning models is their lack of explainability. The internal processes of deep learning models are typically difficult to express. To address this issue, Nagashima, Morita, and Takeuchi (2021) have explored curiosity models aligned with cognitive modeling approaches. By employing cognitive models, it becomes possible to trace processes in a manner comparable to human internal processes.

In their study, adaptive control of thought-rational (ACT-R) was utilized as a cognitive architecture to implement a model of curiosity. ACT-R allocates individual cognitive functions to basic units called modules (Fodor, 1983). These modules in ACT-R correspond to brain regions, and their behavior is being validated through brain measurements using functional magnetic resonance imaging (fMRI) (Anderson, 2005). Furthermore, the mapping between the modules' domains is based on neuroscientific findings (Stocco et al., 2021).

The model developed by Nagashima et al. (2021) specifically utilizes ACT-R's characteristic symbolic processes to represent "pattern discovery," which is humans' ability to identify, combine, and utilize patterns of causal relation-

ships (Baron-Cohen, 2020). This emphasis on pattern discovery aligns with the aforementioned mathematical models (Friston, 2010; Schmidhuber, 2010), which posit that curiosity arises from discrepancies between perceptions of the external world and predictions derived from experience. These differences from predictions generate surprise (curiosity), some of which induce an emotional response such as enjoyment (Koster, 2013; Schmidhuber, 2010).

The curiosity represented in Nagashima et al. (2021)'s model was based on assumed correspondence between pattern discovery and pattern matching in ACT-R's symbolic process. Their model employs curiosity to perform a continuous maze task. The task is implemented by combining the continuation of production and the stopping production at the beginning of the task. This conflict resolution is expressed by ACT-R's utility module and production compilation module. Specifically, they utilized a utility module to assign positive rewards to continuation productions when production compilation occurred. This indicates that when a new production rule is created through compilation, the model experiences a sense of "fun." Conversely, when no compilation occurs, the continuation production receives negative rewards, leading the model to feel "bored." Ultimately, the model terminates the task.

Trait Inference on Artificial Agents

Research on trait inference by humans has accumulated extensively in social psychology. This field relates to studies on anthropomorphism towards artifacts (Nass & Moon, 2000). Within the field of human-agent interaction (HAI) (Laban, 2021), studies on attribution toward artificial agents have been conducted based on previous research on animacy perception Heider and Simmel (1944). However, many studies have manipulated external factors like appearance and behavior without directly manipulating the internal factors underlying agent behavior generation(Van Pinxteren, Pluymaekers, & Lemmink, 2020).

In a few exceptional studies, for instance, Rato, Couto, and Prada (2021) evaluated how agents' traits adapt to different contexts. In this study, participants were presented with the behavior of agents placed in a three-dimensional virtual space. Some agents acted in response to the context of their environment, while others acted randomly. Participants then assessed the traits of the models using motivation-related questionnaires. Similarly, Walker, Weatherwax, Allchin, Takayama, and Cakmak (2020) implemented a robot with a curiosity-based internal model and had humans evaluate whether it exhibited intelligence. They recorded the robot's behavior on videos and presented it to participants online, using surveys to assess the robot's characteristics. The survey utilized the "Perceived Intelligence" indicator from the Godspeed questionnaire (Bartneck et al., 2023), which evaluates characteristics based on an agent's behavior.

The above studies suggest that humans can partially infer the traits of agents from the behavior of machine learning or computational models. However, the correspondence
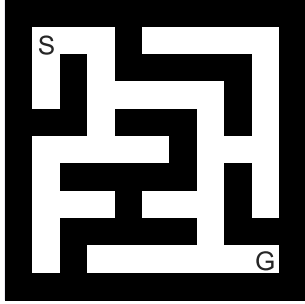
Figure 1: Maze Environment.

between the internal models addressed in these studies and human cognitive functions is not clear. Therefore, in this study, we adopt the cognitive model of curiosity developed by Nagashima et al. (2021) and examine the relationship between the model's predictions and participants' attributions.

Method

Participants

Participants were recruited via a Japanese crowdsourcing website (Lancers.jp) in February 2024. Ninety-five participants were the target of the analysis, excluding incomplete data from a total of 100 participants.

Materials

We prepared a total of three cognitive models of curiosity that differ in the level of thinking applied to the task as experimental factors. Below, we present the simulation tasks conducted by the models, followed by an overview of the three cognitive models corresponding to the experimental conditions established in this study. At the end of this section, we describe the environment in which the cognitive models are presented to the participants.

Simulation Task In the material to the participants, the model explores a set of maps based on curiosity. Figure 1 is an example of a maze map explored by the model. The size of the map was set to 9×9, the widest size among those covered by previous studies.

These environments are represented as topological maps in which the corner points of the maze are nodes, and the connections (paths) between nodes are represented as declarative chunks.

During the task execution, the model discovers the memorized paths (consisting of two consecutive turns and their directions) stored in the declarative module by matching them with patterns embedded in production rules. Each process of the model moving from the start position to the goal position or reaching the time limit (180 seconds) constitutes one round. Multiple rounds on the same map continue until either the overall time limit for the task (3600 seconds) is reached or the model becomes "bored" with the task. The model's enjoyment and boredom are represented based on Nagashima et

al. (2021)'s curiosity cognitive model, as described in related research.

Experimental Conditions To investigate the models' curiosity in the described setup, we varied the agent's traits based on their level of thinking in environmental exploration strategies. Models were categorized by the extent of pattern matching, with higher levels frequently accessing the declarative module for task completion (needing more cognitive load). Lower-level models, however, performed tasks without utilizing this strategy (requiring less cognitive load). In other words, the higher level models are more related to "cognitive curiosity" while the lower level model is more related to "perceptual curiosity" as discussed by Malone (1981). Below is an overview of each model, with each repeating the outlined processes.

1. Random Model:
   The model randomly determines one of the four directions (east, west, south, north). Then, it queries the declarative module to check if it can move in that direction. If movement is possible, it proceeds in that direction.

2. Stochastic DFS Model (DFS):
   The model utilizes stochastic depth-first search for exploring the environment. Similar to the Random Model, it randomly selects the direction of movement and checks its viability. If movement is possible, the model proceeds accordingly and stacks the moved path. In case of encountering a dead-end, it backtracks to return along the previous path.

3. Stochastic DFS + IBL Model (DFS+IBL):
   The model combines stochastic DFS with IBL (instance-based learning) (Gonzalez, Lerch, & Lebiere, 2003). IBL is a learning approach that utilizes memories to solve current tasks. The model behaves essentially like the DFS model. When the model reaches the goal, it stores the path to the goal as declarative knowledge with the "correct" label. The model then utilizes an IBL strategy to approach the goal by recalling these paths. The model explores the environment by switching between these two strategies.

The models are assumed to have increasing levels of thinking complexity from 1 to 3. The random model takes random actions, providing fewer opportunities for deliberation on the next move, hence considered to have a lower level of thinking. On the other hand, DFS+IBL incorporates both the strategy of DFS and the IBL strategy, allowing for more opportunities for deliberative memory retrieval, thus being considered to have a higher level.

As the experimental material, we randomly sampled 10 runs of the model for each of the three conditions. Table 1 shows the statistical measures of the model runs. "Round Num" indicates the number of rounds, "Total Time" denotes the time (in seconds), and "Goal Rate" represents the percentage of times the model reached the goal.

Table 1: Statistics of presented stimuli.

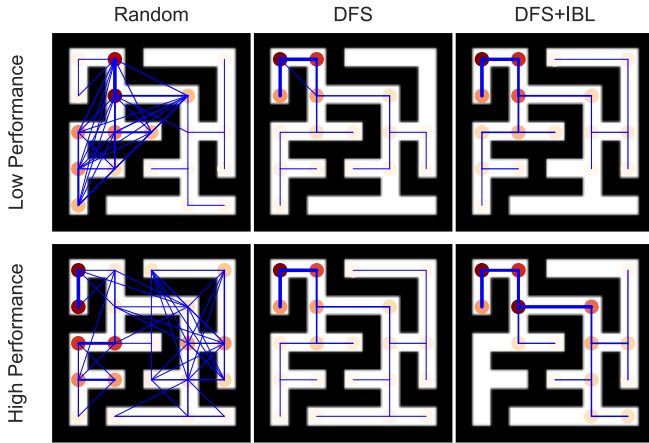|	Round Num				Total Time				Goal Rate			
	Mean	SE	Min	Max	Mean	SE	Min	Max	Mean	SE	Min	Max
Random	33.8	2.62	23	46	3586.2	13.81	3461.9	3600.0	0.56	0.05	0.26	0.76
DFS	9.3	0.98	4	15	1557.5	501.44	665.9	2499.6	0.16	0.03	0.00	0.27
DFS+IBL	6.6	0.65	3	10	980.7	116.58	329.7	1440.0	0.34	0.11	0.00	1.00

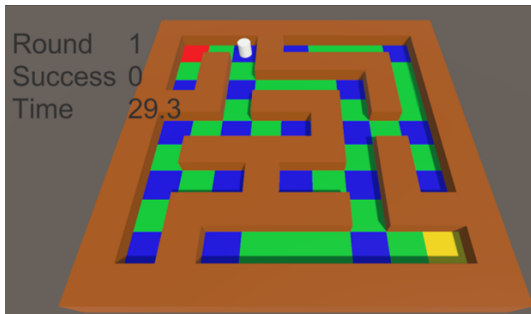Figure 2: Trajectories of the model runs.

Figure 3: Simulator.

Figure 2 illustrates the trajectories of model movements for runs with the lowest and highest goal rates for each model. The color of each circle represents the visit frequency, with darker red indicating more visits by the model. Additionally, the thickness of each line represents the frequency of paths chosen by the model. It should be noted that diagonal or wall-penetrating movements in the random model result from the compilation of multiple movement production rules.

Website for Agent Evaluation Following the previous study (Walker et al., 2020), we recorded the agents' behavior on videos and created a website for participants to present the movie and conduct subjective ratings.

The movies demonstrated the agent's movement on a visualizer created from the model log and Unity. Figure 3 shows this visualizer, which also displays the number of rounds (Round), number of goals achieved (Success), and current

time (Time) to make participants understand the current status of the model. A white cylinder represents the agent in an environment mirroring the map depicted in Figure 1, with start (red), goal (yellow), and corner (blue) interconnected by grid squares. The agent moves between these squares.

The visualizer represents the model's decision-making process between grid squares, with movements synchronized to ACT-R simulation time. We expected participants to form varied impressions of each model feature. Movement speed was kept constant for each model, reflecting ACT-R's simulation without physical body movements.

The recorded movies were uploaded to YouTube, and their links were embedded in the website. During the experiment, participants were able to freely manipulate the controls of the YouTube movies, such as fast-forwarding, changing scenes, and stopping playback. In addition to these videos, the website provided participants with information summarizing the results of the stimuli as presented in Table 1.

Based on the provided information, participants answered a questionnaire regarding the traits of the agents. The questionnaire utilized the "Perceived Intelligence" indicator from Godspeed(Bartneck et al., 2023), similar to previous research (Walker et al., 2020). Additionally, the Five-Dimensional Curiosity Scale (5DS) (Kashdan et al., 2018) was used to investigate curiosity types in detail. The 5DS classifies five types of curiosity, from which "Joyous Exploration" and "Deprivation Sensitivity" were selected for relevance to the simulation task. The former relates to the exploration of things for joy and positive experiences. The latter is associated with exploring intellectual matters such as problem-solving and filling knowledge gaps. We considered that these indicators correspond to the sensory and cognitive aspects of curiosity as outlined by Malone (1981).

Table 2 presents the questionnaires used in this study. As can be seen in the table, each index in the 5DS consists of five questions. Perceived intelligence taken from Godspeed also contains five pairs of antonyms related to the presence or absence of intelligence. All the questionnaires were rated on five-point scales. Those questionnaires were translated into the Japanese language. During this modification process, we also changed pronouns from the original first-person one to the third-person one (from "I" to "He") to enable participants to rate the behavior of the agents.

Procedure

The participants recruited from the crowdsourcing site, Lancers, viewed the instructions and then began the experi-

Table 2: Questionnaire.

	Joyous exploration
1	He views challenging situations as an opportunity to grow and learn.
2	He is always looking for experiences that challenge how he thinks about himself and the world.
3	He seeks out situations where it is likely that he will have to think in depth about something.
4	He enjoys learning about subjects that are unfamiliar to him.
5	He finds it fascinating to learn new information.
	Deprivation sensitivity
1	Thinking about solutions to difficult conceptual problems can keep him awake at night.
2	He can spend hours on a single problem because he just can't rest without knowing the answer.
3	He feels frustrated if he can't figure out the solution to a problem, so he works even harder to solve it.
4	He works relentlessly at problems that he feels must be solved.
5	It frustrates him not having all the information he needs.
	Perceived Intelligence
1	Incompetent - Competent
2	Ignorant - Knowledgeable
3	Irresponsible - Responsible
4	Unintelligent - Intelligent
5	Foolish - Sensible

Table 3: Number of participants for each simulation.

	1	2	3	4	5	6	7	8	9	10
Random	13	11	13	8	7	9	6	9	9	10
DFS	9	9	7	9	9	12	9	13	6	12
DFS+IBL	8	10	7	6	11	13	10	6	13	11

Figure 4: Result of the questionnaire.

ment. The following is the experimental procedure.

1. Register the participation on the Lancers' request screen

2. Read the instructional page

3. Repeat the following operations on the task screen three times each corresponding to one of the models of the three levels of thinking

   (a) Observe the movie and read the information about the model run.

   (b) Fill out the questionnaire

The participants were informed that the task involved inferring the traits of three agents based on the presented materials (the movie and the information about the run). They were also instructed to observe the information for each agent for at least three minutes. The webpage displaying each model run was controlled by JavaScript and did not display a 15-question questionnaire for trait rating until three minutes had passed.

In the instructions, they received an overview of the agent's task, which involved repeated maze-solving. They were also informed about the completion conditions (being "bored" and the time limit) of the simulation task. Additionally, they were informed that the agent had the learning ability to reduce thinking time for solving the maze. Furthermore, they were provided with explanations about the presented materials (simulation movies, total round count, goal rate, and total thinking time). The functionality of the movie player was explicitly noted, allowing participants to freely fast forward and change scenes. Finally, participants were notified of dummy questions for careful answering (though not included).

Once participants self-assessed their understanding of the instructions as sufficient, they proceeded to the task screen. The presentation order of the three models was randomized. On the task screen for each model, participants were presented with a selection from 10 simulations chosen randomly. Thus, the number of participants varies in each simulation. Table 3 indicates the frequency of selection for each simulation.
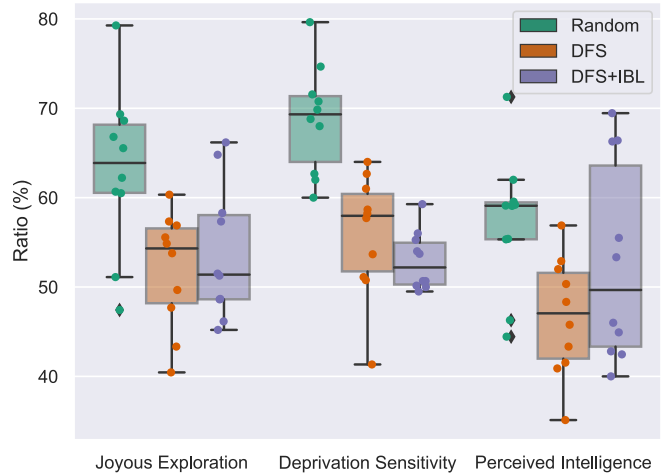
Results and Discussion

Effect of Internal Process on Trait Inference

To examine the influence of the model's internal processes on participants' trait inferences, we calculated the average scores for two indicators of the 5DS and one indicator of the Godspeed for each model. Following prior research (Kashdan et al., 2018), these scores were converted to proportions relative to the maximum value (5) of the questionnaire. After calculating the average scores for each index and each participant, we aggregated the scores for each model run. The sample size for this aggregation (calculating the averages) corresponds to the numbers in Table 3. Thus, we used model runs as the unit of analysis ($n = 10$ for each model).

Figure 4 depicts the results of the questionnaire for each model. The color of each box represents the level of thinking about the models. Each item on the x-axis indicates the indicators of 5DS (Joyous Exploration and Deprivation Sensitivity) and Godspeed's Perceived Intelligence. The y-axis represents the mean of the evaluation scores obtained for each questionnaire item, averaged across model executions.

We examined the differences between models using analysis of variance for each indicator shown in Figure 4. As a result, significant differences between models were found for all indicators (Joyous Exploration: $F(2,9) = 8.02$, $p < .01$; Deprivation Sensitivity: $F(2,9) = 21.77$, $p < .01$; Perceived Intelligence: $F(2,9) = 3.87$, $p < .05$). The Holm method for multiple comparisons revealed that the Random model was significantly rated higher than other models in two curiosity-related indicators ($p < .05$). Regarding intelligence-related indicators, a significant difference was observed only be-

Table 4: Correlations between the presented stimulus and the mean of the indicators corresponding to the models in Table 1. All Models summarizes all models ($n = 30$). Others correspond to individual models ($n = 10$).

	All Models			Random			DFS			DFS+IBL		
	Round	Time	Goal	Round	Time	Goal	Round	Time	Goal	Round	Time	Goal
Joyous	0.57**	0.44	0.65**	0.40	-0.24	0.39	-0.06	-0.12	0.11	-0.16	-0.50	0.79**
Deprivation	0.79**	0.72**	0.43	0.58	0.36	0.30	-0.26	-0.28	-0.31	-0.39	-0.45	0.29
Intelligence	0.40	0.21	0.84**	0.63	-0.11	0.68	-0.08	-0.16	0.76	-0.27	-0.64	0.87**

$** \; p < 0.01$

tween the Random model and the DFS model ($p < .05$), with no difference observed otherwise.

Correlation between Model Behavior and Trait Inference

As described above, participants perceived the Random model as the most motivating and rated it as more intelligent compared to the DFS model, potentially inferring traits based on long task persistence and goal rates. As shown in Table 1, the Random model spent more time on tasks and achieved a higher goal rate compared to other models.

To examine the relationship between such external behaviors and trait inference, we calculated correlation coefficients between behavioral indicators in Table 1 and evaluation scores in Figure 4 (see Table 4) to explore the relationship between external behaviors and trait inference. Table 4 presents results for each model individually and combined for all models. Correlations involving all models showed a relationship between the round and the two curiosity indices, the goal rate and Joyous Exploration, and the total time and Deprivation Sensitivity. These findings suggest participants perceived higher intelligence in models with a higher goal rate and associated greater curiosity with longer task persistence.

However, the overall results do not clearly differentiate based on internal algorithms. Correlations between goal rate, intelligence, and Joyous Exploration were only observed in the DFS+IBL model, reflecting its diverse behaviors in this model. In fact, Table 1 highlights a higher variance in goal rate for the DFS+IBL model compared to others.

Conclusion

This study explored the utility of cognitive models for investigating human trait inference, representing a novel application of such models. Participants inferred traits from the behavior of a cognitive model implemented using ACT-R. Results indicated that participants' trait inferences were primarily influenced by the model's behavior rather than its internal processes. Participants demonstrated heightened curiosity towards models with increased rounds and longer task durations while attributing higher intelligence to models with higher rates of goal attainment. Interestingly, participants tended to rate models displaying more moves and random behavior more favorably than those demonstrating deliberative thinking and moderate curiosity exploration.

The tendency to value random behavior can be interpreted

from research on the perception of intelligence from randomness, as highlighted in the introduction (Fyfe et al., 2008). Humans find it difficult to distinguish between random algorithms and behavior based on derivative intelligence from external behavior. Furthermore, random behavior has been suggested to have adaptive value in contexts such as creative thinking (Cropley, 2006; Runco & Jaeger, 2012), so attributing high intelligence to random behavior cannot be categorically dismissed. Additionally, it cannot be denied that the cultural background of the participants targeted in this study may have influenced the results. It has been reported that in Japanese work culture, long working hours are considered virtuous (Ono, 2018).

Regardless of the reasons, the findings of this study align with observations on attribution errors emphasized in social psychology. The significance of this research lies in its utilization of cognitive models for studying trait inference. With a focus on internal processes, cognitive models offer the advantage of traceability in human inference on the model's internal process compared to machine learning agents. Through this analysis, cognitive models serve as tools for revealing biases in human trait inference. In the future, further advancement of this approach is necessary. The main finding is that participants associated curiosity and intelligence with models displaying high randomness. Contrary to this main finding, the DFS+IBL model, representing the highest level of thinking, demonstrated a correlation between its goal rate and perceived intelligence, suggesting that individuals perceive varying intelligence levels in different runs of the same model. This result suggests that participants found a different type of intelligence in the DFS+IBL model compared to the random model. Future studies are needed to understand the conditions shaping human perceptions of such deliberative intelligence.

Moreover, to delve into the intricacies of human trait inferences, enhancements in the information and environment provided to participants are crucial. While this study involved presenting model behaviors over prolonged periods, it is unrealistic to anticipate scrutiny from participants recruited through crowdsourcing. Hence, for humans to more effectively estimate models in the future, ongoing improvements in the tasks undertaken by the models, provision of information regarding their internal workings, and environmental examination are necessary.

References

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(3), 313–341.

Aubret, A., Matignon, L., & Hassas, S. (2019). A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*.

Baron-Cohen, S. (2020). *The pattern seekers: How autism drives human invention*. New York: Basic Books.

Bartneck, C., Cochrane, T., Nokes, R., Chase, G., Chen, X., Cochrane, T., . . . Adams, B. (2023). Godspeed questionnaire series: Translations and usage.

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, *18*(3), 391–404.

Fodor, J. A. (1983). *The modularity of mind*. MIT Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Fyfe, S., Williams, C., Mason, O. J., & Pickup, G. J. (2008). Apophenia, theory of mind and schizotypy: Perceiving meaning and intentionality in randomness. *Cortex*, *44*(10), 1316-1325.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Kashdan, T. B., Stiksma, M. C., Disabato, D. J., McKnight, P. E., Bekier, J., Kaji, J., & Lazarus, R. (2018). The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, *73*, 130–149.

Koster, R. (2013). *Theory of fun for game design*. Sebastopol: O'Reilly Media.

Laban, G. (2021). Perceptions of anthropomorphism in a chatbot dialogue: The role of animacy and intelligence. In *Proceedings of the 9th international conference on human-agent interaction* (pp. 305–310).

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, *5*(4), 333–369.

Nagashima, K., Morita, J., & Takeuchi, Y. (2021). Curiosity as pattern matching: Simulating the effects of intrinsic rewards on the levels of processing. In *Proceedings of the 19th international conference on cognitive modelling* (p. 197-203).

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81–103.

Ono, H. (2018). Why do the japanese work long hours? sociological perspectives on long working hours in japan. *Japan Labor Issues*, *2*(5), 35-49.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Rato, D., Couto, M., & Prada, R. (2021). Fitting the room: Social motivations for context-aware agents. In *Proceedings of the 9th international conference on human-agent interaction* (pp. 39–46).

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247.

Stangor, C. G., & Walinga, J. (2014). *Introduction to psychology - 1st canadian edition*. Victoria: BCcampus.

Stocco, A., Sibert, C., Steine-Hanson, Z., Koh, N., Laird, J. E., Lebiere, C. J., & Rosenbloom, P. (2021). Analysis of the human connectome data supports the notion of a "Common Model of Cognition" for human and human-like intelligence across domains. *NeuroImage*, *235*, 118035.

Van Pinxteren, M. M., Pluymaekers, M., & Lemmink, J. G. (2020). Human-like communication in conversational agents: A literature review and research agenda. *Journal of Service Management*, *31*(2), 203–225.

Walker, N., Weatherwax, K., Allchin, J., Takayama, L., & Cakmak, M. (2020). Human perceptions of a curious robot that performs off-task actions. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (pp. 529–538).

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*(2), 237.

Exploring an Approach for Phonological Awareness Estimation Employing Personalized Cognitive Models and Audio Filters

Jumpei Nishikawa (nishikawa.jumpei.16@shizuoka.ac.jp)
Junya Morita (j-morita@inf.shizuoka.ac.jp)
Shizuoka University,
3-5-1, Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011, Japan

Abstract

Language development is supported by phonological awareness, which is related to attention to phonological aspects of spoken language. We aim to develop a system that supports phonological awareness formation using cognitive models. Estimating the state of a user's phonological awareness is a kind of identification of the user's "auditory filter." This paper reports on an experiment with typically developed native speakers by setting up an audio filter that is applied to the system's output sound. The user's phonological awareness is estimated as a relative preference for two computational models presented by the system. Using the system with audio filters, we test the hypothesis that there is a difference in participants' selection behavior depending on the characteristics of the model under the application of the audio filter. The results of the experiment showed that there was a difference in selection probability between models with different degrees of sound confusion, with and without the application of a specific audio filter.

Keywords: ACT-R, Cognitive modeling, Phonological awareness, Personalized model

Introduction

Children (or second-language learners) face various difficulties during language acquisition. A prominent example is the segmentation of phonemes. In the early stages of language development, children perceive speech sounds as continuous but can gradually segment them into smaller units (Carroll, Snowling, Stevenson, & Hulme, 2003). In the process, sound can be segmented into various units (symbols), such as syllables and morae. As learners advance, they converge on a system of processing a series of units as defined by their native language (e.g., mora in Japanese; Kubozono, 1989).

In the fields of developmental psychology and speech–language pathology, one of the abilities supporting this development is phonological awareness, which involves paying attention to phonological aspects of speech, such as phonemes and rhythm (Stahl & Murray, 1994). Some speech errors that occur during language development are attributed to a poorly formed phonological awareness of that particular language (Dynia, Bean, Justice, & Kaderavek, 2019; Kobayashi, 2018; Smith Gabig, 2010). In children with autism spectrum disorder (ASD), an overall delay in phoneme acquisition and a partial inability to use some phonemes may occur (Grandin & Panek, 2013; Mugitani et al., 2019).

Computational modeling is effective for understanding and predicting human internal processes, such as phonological awareness, that cannot be directly observed. Based on this idea, a system to support the formation of phonological awareness using computational models has been developed (Nishikawa & Morita, 2022a, 2022b). In this study, we evaluate a method for estimating individual phonological awareness using computational models to realize this long-term goal. In the experiment where participants virtually faced difficulties in language learning, we recruited typically developed native speakers as participants. This paper reports the results of the experiment.

System

In a previous study (Nishikawa & Morita, 2022a), a model of phonological awareness using the cognitive architecture ACT-R (Anderson, 2007) was implemented. This model maps the general memory retrieval mechanism of ACT-R to phonological awareness and represents errors during *Shiritori* to simulate immature phonological awareness.

Shiritori is a Japanese word game. This game involves players taking turns uttering a word (noun); the word must begin with the mora that the previous word ended with. For example, after a player answers *"ri-n-go"* (meaning apple), the next player continues with *"go-ma"* (meaning sesame seeds). Specifically, the model's memory of morae[1] has a similarity value. Shiritori errors due to "mistaking similar sounds" can be represented as false retrievals affected by the similarity set between morae when retrieving words cued by morae. In addition, the model can be varied by adjusting the parameters of ACT-R. For example, two of the parameters are manipulated: the method of computing the similarity between morae and the value of the coefficient P, which corresponds to the magnitude of the effect of the similarity. These parameter adjustments allow some models to accommodate specific errors found in children, such as "consonant deletion" (Oishi, 2016; Grandin & Panek, 2013).

Figure 1 shows an overview of the Phonological Awareness Formation Support System (Nishikawa & Morita, 2022b). The system includes several variants of phonological awareness models based on previous research. The system

---

[1] one of the units of sound. It is defined by the duration time (Port, Dalby, & O'Dell, 1987). In the Japanese language, this is considered the basic unit.
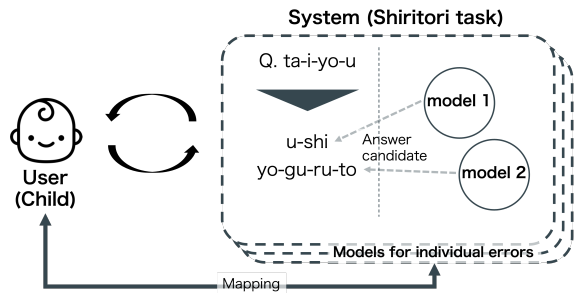
Figure 1: System overview. Phonological awareness is estimated as the user plays choice-based Shiritori. The words are for example.

provides a "choice-based Shiritori," where players select appropriate words from a set of alternatives presented by the model. Estimation of phonological awareness using the system consists of multiple sessions of repeated choice-based Shiritori. In each session, the system has two models to produce alternatives. By changing the models in each session, the system identifies a model that has the corresponding characteristics to the user's phonological awareness. For example, a learner with a phonological processing difficulty, such as consonant deletion, may be unaware of incorrect candidates suggested by a model with characteristics similar to his or her own.

Method

The concept for one experiment session is shown in Figure 2. Estimating the phonological awareness of a user can be described as a task to identify the user's auditory filter for speech. Thus, we believe that the source of individual differences in language learning can be attributed to personal filters that convert physical sounds into segments, enabling individuals to perceive units.

In examining the validity of this task setting, it is difficult to control for the user's filter in experiments with real users (i.e., children and second language learners). In light of this issue, the present study uses audio filters set by the authors. Through an experiment with four conditions of models, which are combinations of the two methods of computing similarity between morae and the existence of audio filters (with vs. without), we make the following assumption: *if the characteristics of the user's phonological awareness match the characteristics of the model, the user will select candidate words that the model incorrectly presents without realizing it.* Based on this assumption, this experiment tests the hypothesis that there is a difference in participants' selection behavior depending on the characteristics of the model under the application of the audio filter, using the frequency with which the participant selects the model as an indicator. In this paper, we focus on the degree of sound confusion in the model (corresponding to P in ACT-R) and the effect of the speech filter on the selection behavior. The experimental setup for this is described below.
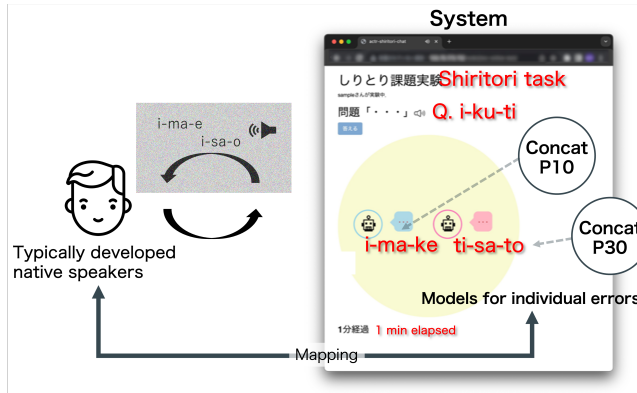
Figure 2: Experimental Concept. This is an example of a one-session of choice-based Shiritori with nonsense words. The output of the system is filtered to simulate the characteristics of individual phonological awareness. The red text is a note by the author, and during the experiment, the words are presented only as audio played by clicking on icons.

Table 1: Experimental conditions (model and audio filter) and number of participants

	Vowel filter	No filter
Concat (lax, strict)	49	47
Ave (lax, strict)	44	45

Participants

In this study, we test the identification of audio filters set by the authors. That is, we set up an audio filter that applies to the system's output speech as a filter owned by the user, and the participants are typically developed Japanese native speakers.

Participants were recruited via the Japanese crowdsourcing platform Lancers[2]. Participants who accessed the recruitment page were distributed to one of the four conditions via the links provided. Table 1 shows the model used, the audio filter, and the number of participants for each experimental condition. The model and filters are described in detail in the following subsections. One hundred and eighty five participants (79 females, five non-respondents, $mean_{age} = 44.2$, $SD_{age} = 10.2$) participated. Table 1 shows the number of participants used in the analysis for each condition [3].

The model used in the experiment

Four models were used in the system throughout the experiment. The parameters of the model were manipulated at two levels (Concat vs. Ave) in the way the mora similarity was calculated[4] and at two levels (10 vs. 30) in the similarity impact P. The model with Concat-type similarity is the one whose behavior has been pointed out to be related to consonant deletion in a previous study (Nishikawa & Morita, 2022a).

The two models, which are presented to the participants simultaneously in each condition in the experiment, differ only in the similarity impact P (i.e., degree of sound confusion, which can also be described as the frequency of proposing incorrect answer candidates[5]). Based on this, we refer to a $P = 10$ model as a *lax model* and a $P = 30$ as a *strict model*. Based on this setting, we can predict that participants will be unable to distinguish between the behavior of two models when the characteristics of the model and the characteristics of the audio filter match and will be able to select a less erroneous model when the characteristics of the model and the characteristics of the audio filter do not match (because they can discriminate errors even in the filtered situation). Therefore, this experiment tests the hypothesis that errors in models that fit the filter are selected more often.

The models in the system are assumed to have a common vocabulary of 2,000 nonsense words generated by combining three morae of 100 different morae[6]. Candidate words proposed by the models in the choice-based Shiritori are output from the system by playing an audio file. We prepared audio files (mp3) of 2,000 words in the model's vocabulary using the text-to-speech service Amazon Polly[7]. In the SSML (Speech Synthesis Markup Language) specified to create the audio files, "x-amazon-pron-kana" was specified in the alphabet tag, and *Katakana*[8] in the ph tag.

Audio filter

The application of an audio filter, called *vowel filter* in this paper, to the system's output (i.e., the model's reading of candidate words) can be achieved by using audio files with incorrect utterances. When creating the audio files, adjustments are made to the correct vocabulary files to prepare the incorrectly uttered vocabulary files. In this experiment, we prepared error patterns in which the initials and endings of words are replaced with vowel-only morae, to

---

[4]For details, see the previous study (Nishikawa & Morita, 2022a). When vectorizing morae for similarity calculations, there is a difference between concatenating (Concat) or averaging (Ave) the phonemes in a mora.

[5]see Nishikawa and Morita (2022a) for details. $P = 10$ is more frequently incorrectly answered.

[6]The mora used to generate the mora excludes "N," which cannot exist at the beginning of a word. The frequency of morae in nonsense words is based on their frequency of appearance in the Japanese dictionary (Amano & Kobayashi, 2008).

[7]https://aws.amazon.com/de/polly/

[8]A kind of Japanese character.

Table 2: Example of an audio file. Words are shown in Japanese characters. Supplementary information written in parentheses in alphabetical characters.

	Original	Incorrect utterances
1	いくち (i-ku-ti)	イクイ (i-ku-i)
2	いまけ (i-ma-ke)	イマエ (i-ma-e)
3	ちさと (ti-sa-to)	イサオ (i-sa-o)
⋯	⋯	⋯

map them to phenomena such as consonant deletion. Table 2 shows an example of a false utterance created by this rule. The correct and incorrectly voiced audio files are shown with the pronunciation of the words in Japanese characters.

Preliminary analysis of results

This system assumes that when the characteristics of the user's phonological awareness match those of the model, the user selects a candidate answer word that stood incorrectly suggested by the model without realizing it. Based on this assumption, this experiment tests the hypothesis that there is a difference in participants' selection behavior depending on the characteristics of the model under the application of the audio filter. In particular, to investigate the effect of sound confusion of the models (strict or lax, corresponding to P in ACT-R), the difference of the patterns of the sound confusion (Concat or Ave), and the effect of the audio filter on the selection behavior, we first summarize the participants' responses by pattern and then aggregate it by these factors.

Table 3 shows the participants' behavior patterns during the Shiritori task. The table's columns correspond to four pattern categories (both models answer correctly, only one answers correctly, only another answers correctly, and both models answer incorrectly). Under these four patterns, participants' behavior can be further divided into two categories, depending on which model is selected (rows in the table). Each value in a cell indicates the total number of selections the participant made in that condition. Cells colored gray indicate cases where the participant answered Shiritori incorrectly. Focusing on the pattern where only one model was correct (the middle two columns of each table), we can see that the correct model is more often chosen. This is the same for all patterns and filter conditions.

To normalize the values in Table 3, which are influenced by the number of participants and the number of Shiritori continuations, we calculated the posterior probability by

$$P(Choice_i|Pattern_j) = \frac{P(Choice_i) \cdot P(Pattern_j|Choice_i)}{P(Pattern_j)}$$
(1)

where $Pattern_j$ is a possible state of the system and takes one of the four values corresponding to the column names in Table 3. The $Choice_i$ is the participant's choice and takes one of two values corresponding to the row names in Table 3. For example, $P(Pattern_j)$, the denominator of the equation 1, is

Table 3: Classification of participant behavior

(a) Concat models–Vowel filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	2556	99	1350	156
User choose strict	2568	40	3213	210

(b) Concat models–No filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	2639	142	252	114
User choose strict	2136	16	3991	211

(c) Ave models–Vowel filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	2154	138	1964	249
User choose strict	2255	76	2568	319

(d) Ave models–No filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	2262	159	370	208
User choose strict	2164	11	3276	318

Table 4: Posterior probability of participant behavior

(a) Concat models–Vowel filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	0.50	0.71	0.30	0.43
User choose strict	0.50	0.29	0.70	0.57

(b) Concat models–No filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	0.55	0.90	0.06	0.35
User choose strict	0.45	0.10	0.94	0.65

(c) Ave models–Vowel filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	0.49	0.64	0.29	0.44
User choose strict	0.51	0.36	0.71	0.56

(d) Ave models–No filter condition

| | Lax model correct | | Lax model wrong | |
	Strict model correct	Strict model wrong	Strict model correct	Strict model wrong
User choose lax	0.51	0.94	0.10	0.40
User choose strict	0.49	0.06	0.90	0.60

the probability obtained by dividing the sum of a column by the total sum of the table. The calculated values are shown in Table 4. The trends of the probabilities are consistent with those of the frequencies (Table 3), showing the selection tendency toward the correct answers regardless of the filter or the model variations.

To visualize the effects of the filter and model variations, we computed the differences in posterior probabilities between the strict and lax models (Figure 3a) and between the Concat and Ave models (Figure 3b). In both cases of Figure 3, positive values mean a higher probability that the models with filtered conditions were selected, while a negative value means a higher probability that the models with no filter conditions were selected. That is, Figure 3a shows the values that are subtracted from the weighted average of Tables 4a and 4c to the weighted average of Tables 4b and 4d, while Figure 3b compares the difference in selection probability for each filter condition for the lax model[9] between the Concat and Ave types. Specifically, the top row of Table 4a minus the top row of Table 4b is shown as right-hatched bars, and the top row of Table 4c minus the top row of Table 4d is shown as dot-hatched bars.

In every model condition (strict vs. lax / Concat vs. Ave), we can find more biased choices in the patterns where only one model was correct (the middle two patterns of each figure). However, the biases are different in the different

directions of the models. By focusing on Figure 3a, we can find clear differences between two bars (orange and blue bars), showing that the wrong models (the strict model in the strict wrong/lax model in the lax wrong) have a bias toward filtered choice (positive value), and the correct models (the strict model in the strict correct/lax model in the lax correct) have a bias toward no filtered choice (negative value). In other words, the model that presented the wrong answer was relatively well selected in the vowel filter condition, while the model that presented the correct answer was relatively well selected in the no filter condition. These results suggest that the audio filter set in the study matches the lax models in the experiment.

In contrast to the clear difference between Strict and lax in part of Figure 3a, the difference between Concat and Ave is difficult to discern in the comparison in Figure 3b. The two bars in Figure 3b come s from the blue bar in Figure 3a. Therefore, the two bars are filtered in when the model is wrong and unfiltered when the opposite model is wrong. Although there is a slight difference, right-hatched bars are larger than dot bars in these patterns, indicating that the Concat model was relatively selected in the vowel filter condition. Previous research (Nishikawa & Morita, 2022a) suggested a link between the Concat model and consonant deletion and that the vowel filter was inspired by consonant deletion, so these results are consistent with the hypothesis and assumption of the experiment.

---

[9]The Strict model selection is a complementary event to the Lax model selection in specific patterns. The present analysis focuses on the Lax model, which presents more wrong choices.

(a) The difference in posterior probabilities between models in vowel filter and no filter condition. These values are the weighted average of Tables 4a and 4c minus the weighted average of Tables 4b and 4d.
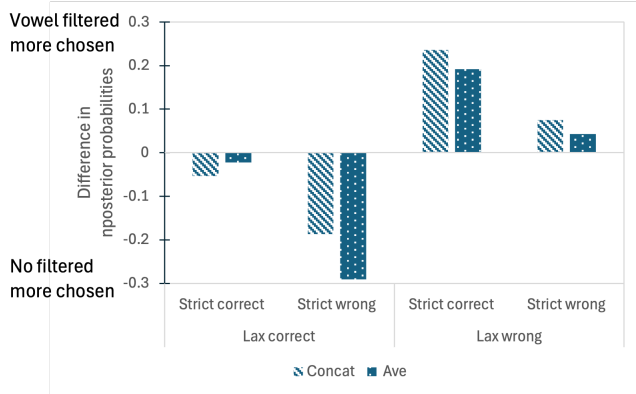
(b) The difference in posterior probabilities between lax models in vowel filter and no filter condition. The right-hatched bars are the top row of Table 4a minus the top row of Table 4b, and the dot-hatched bars are the top row of Table 4c minus the top row of Table 4d

Figure 3: Differences in posterior probabilities between models

Conclusion

This paper presented an experiment to evaluate the feasibility of a phonological awareness formation support system using a personalized cognitive model, with typically developed native speakers as participants by applying an audio filter. Based on the hypothesis that there is a difference in participants' selection behavior depending on the characteristics of the model under the application of the audio filter, the results of the experiment showed that there was a difference in the selection probability between strict and lax models depending on the filter conditions. We also found that the Concat model was more chosen when the participants were applied by the vowel filters. These results are consistent with the hypothesis. That is, it provides a clue toward the realization of a method for estimating an individual's phonological awareness using a computational model.

A more detailed data analysis is needed to obtain a clear indicator that can be used as a criterion for the phonological awareness estimation method. In this paper, we only show that there is a difference in selection probability, but we believe that a more detailed analysis of how the difference appears will enable estimation based on characteristics of the model that can be mapped to phenomena such as consonant deletion, as pointed out in previous studies, and will bring us closer to realizing the system. It is also necessary to examine the audio filter settings. By creating a filter based on the knowledge gained from research that reproduces the characteristics of children with difficulties (the unique perceptions of autism spectrum disorder) (Qin, Nagai, Kumagaya, Ayaya, & Asada, 2014), it will be possible to verify that the system corresponds more closely to real-life difficulties. After brushing up on the system based on the results of these experiments, we will conduct system evaluation experiments with language learners, who are the original target of the system.

References

Amano, S., & Kobayashi, T. (2008). *Kihongo database : gogibetsu tangoshimmitsudo.* Gakken plus.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Carroll, J., Snowling, M., Stevenson, J., & Hulme, C. (2003). The development of phonological awareness in preschool children. *Developmental Psychology*, *39*(5), 913–923.

Dynia, J. M., Bean, A., Justice, L. M., & Kaderavek, J. N. (2019). Phonological awareness emergence in preschool children with autism spectrum disorder. *Autism & Developmental Language Impairments*, *4*, 2396941518822453.

Grandin, T., & Panek, R. (2013). *The autistic brain: Thinking across the spectrum.* Houghton Mifflin Harcourt.

Kobayashi, H. (2018). *Oninishiki no keisei to kotoba no hattatsu -"kotoba ga osoi" wo kangaeru-.* Kodama shuppan.

Kubozono, H. (1989). The mora and syllable structure in japanese: Evidence from speech errors. *Language and Speech*, *32*(3), 249–278.

Mugitani, R., Homae, F., Hiroya, S., Satou, Y., Shirose, A., Tanaka, A., … Tachiiri, H. (2019). *Kodomo no onsei* (No. 21). Corona Publishing Co., Ltd.

Nishikawa, J., & Morita, J. (2022a). Cognitive model of phonological awareness focusing on errors and formation process through shiritori. *Advanced Robotics*, *36*(5-6), 318–331.

Nishikawa, J., & Morita, J. (2022b). Estimating phonological awareness with interactive cognitive models: Feasibility study manipulating participants' auditory characteristics. In *Proceedings of iccm 2022, 20th international conference on cognitive modelling* (pp. 203–209).

Oishi, N. (2016). Hyoka [evaluation]. In H. Ishida & I. Ishizaka (Eds.), *Gengochokakushi no tameno gengohattatsushogaigaku [developmental language*

disorder for speech-language pathologists] (2nd ed., pp. 77–117). Tokyo: Ishiyaku Publishers, Inc. (Japanese)

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in japanese. *The Journal of the Acoustical Society of America*, *81*(5), 1574–1585.

Qin, S., Nagai, Y., Kumagaya, S., Ayaya, S., & Asada, M. (2014). Autism simulator employing augmented reality: A prototype. In *4th international conference on development and learning and on epigenetic robotics* (pp. 155–156).

Smith Gabig, C. (2010). Phonological awareness and word recognition in reading by children with autism. *Communication Disorders Quarterly*, *31*(2), 67–85.

Stahl, S. A., & Murray, B. A. (1994). Defining phonological awareness and its relationship to early reading. *Journal of educational Psychology*, *86*(2), 221–234.

A Comparison of Frequency Effects in Two Attitude Retrieval Models

Mark Orr[1] (morr@ihmc.org), Christian Lebiere[2] (cl@cmu.edu)
Don Morrison[2] (dfm2@cmu.edu), Peter Pirolli[1] (ppirolli@ihmc.org)

[1]Institute for Human and Machine Cognition
Pensacola, FL 32502 USA

[2]Department of Psychology, Carnegie Mellon Univ.
Pittsburgh, PA 15223 USA

Abstract

The psychological literature has put forth several auto-associative memory models of attitude formation and change. The status of frequency effects in such models is not well understood. We compare frequency effects in auto-associative memory models of attitudes to the well-established frequency effects found in the ACT-R cognitive architecture. We found striking differences between the model classes, but only under some conditions. We discuss future directions that might stem from this provisional work.

Keywords: attitudes, cognitive modeling, neural networks, memory, dynamical systems

The Problem

Attitude learning is divided into two camps. In one, we have memory processes as a central theoretical component for understanding how attitudes are formed and retrieved. These typically concern memory for the valence towards an attitude object. Although typically not formalized, a running debate in the social psychological literature stems from differentiating or not between simple associative learning or propositional learning in attitudes. This literature is rich in terms of evidence on learning (see Corneille & Stahl, 2019, for examples).

In the other camp, what we will call schema-like memory models, the primary interest is in attitudinal structure (Eagly & Chaiken, 1993). Recent work in this area uses auto-associative memory models to represent not only structure but also as models of attitudinal memory retrieval (e.g., Dalege et al., 2016, 2018). In this work, sets of beliefs are transformed from survey data into a network of associations (e.g., correlations) and modeled using Hopfield-like or Ising-like models. Learning is not well studied in such models. In its place are notions of persuasion: under what conditions will a person stray from their typical attitude retrieval pattern.

In short, little overlap exists between these two literatures. We attempt a kind of reconciliation between the two by studying attitude learning in the auto-associative memory case. Much is know about learning in auto-associative memory systems (e.g., Hopfield, 1982; Hertz et al., 1991). So, we thought it would be useful to directly compare learning in the auto-associative case to learning in an empirically-grounded cognitive architecture. For our comparison, we chose the currently prominent Causal Attitude Network (CAN) model (from social psychology, Dalege et al. (2016, 2018)) to the

ACT-R cognitive architecture (Anderson et al., 2004). Our comparison method, thus, affords the following features: (i) it will ground the findings in human memory systems via ACT-R and (ii) it addresses learning in the structural approach to attitudes.

Design

Across two studies, we compared directly an ACT-R declarative memory model to the CAN attitude model. By directly, we mean that the input data, the model task and the analysis methods were identical. There were some differences in computed measures, but the semantics between them were close.

The Causal Attitude Network Model

The CAN model (Dalege et al., 2016; Dalege & van der Maas, 2020; Dalege et al., 2018) was motivated by the need to provide a dynamic attitude memory retrieval system that exhibits sensitivity to cues in the social environment. Virtually all theoretical work on the CAN model uses fixed, predetermined weights for its network (see below for the formal specification of the system). The CAN model literature references Hebbian learning as a potential candidate for learning attitudes, yet there have been no studies to date that implement learning. The heart of this theoretical work focuses on dynamical retrieval methods that are derived from Ising-like or discrete Hopfield models. The technical details of how a typical CAN model is implemented are as follows–we start with key definitions:

- There is a graph $G = G(V, E)$ consisting of a collection of beliefs (the set of n vertices V) and relations between them (the set of weighted edges E).

- The state of vertex $i \in V$ is $x_i \in K_i$ where K_i is the state set for that vertex.

- For all i we have $K_i \in \{0, 1\}$

- The system state is $x = (x_1, x_2, \ldots, x_n)$.

- The system global energy H is defined using all $i \in V$ by $H(x) = -\sum_{i \in G} \tau_i x_i - \sum_{j \in N_G(i)} w_{ij} x_i x_j$ where $N_G(i) \subset V$ is the set of neighbors of i in G, *not* including i, w_{ij} is the weight of the edge $\{j, i\}$ and τ_i is the baseline parameter for vertex i. Assume that $w_{ij} = w_{ji}$.

- For $i \in V$ let $\sigma_i \colon \prod_{i=1}^{n} K_i \longrightarrow \mathbb{R}$ be the function defined by $\sigma_i(x) = H(x)^c - H(x)^o$ where c and o are the current and opposite state of vertex i.

- For each vertex i we define its vertex function as $\phi_i(x) = 1/(1 + e^{-\sigma_i(x)/t})$ where t is the temperature of the system; this defines the probability that at any point in time a vertex i will flip to its opposite state: $P(c \longrightarrow o) = \phi_i(x)$.

A typical instance of CAN is a discrete-time, asynchronous simulation. For each time step: (i) select a vertex i, (ii) compute $P(c \longrightarrow o) = \phi_i(x)$ and (iii) use $P(c \longrightarrow o)$ directly to decide if vertex i will change its state. Another common implementation is to draw n samples of the system state x from the Gibbs probability distribution. This is computed as: (i) compute the Gibbs probability distribution of all system states x_i such that each is $P(x = x_i) = e^{-H(x_i)}/Z$ where $H(x_i) = -\tau_i x_i - \sum_{j \in N_G(i)} w_{ij} x_i x_j$ and $Z = \sum_X e^{-H(x)}$, (ii) sample from this distribution n times. In our CAN simulations below, we leverage the latter.

ACT-R Declarative Memory

For this article, we develop a comparison to the CAN model using the declarative memory module of the ACT-R cognitive architecture implemented in the PyACTUp Python package[1].

Declarative memory is a module in the ACT-R cognitive architecture comprised of discrete data objects called *chunks*. Each chunk contains a number l of slots which contain attribute-value pairs. The attribute is the slot name and the value is the slot content. Access to this symbolic content is controlled by a subsymbolic quantity called activation, which reflects the characteristics of the knowledge including its history and semantics. The activation calculus determining declarative memory access works as follows:

- The activation A of a chunk is defined as: $A_i = B_i + \varepsilon_i + P_i + S_i$ where B_i is the base level activation, ε_i is stochastic noise, P_i is the partial matching correction, and S_i is the spreading activation . The latter term was not used in the work presented here.

- The base level activation B_i is defined as: $B_i = ln(\sum_i t_{ij}^{-d})$ where t is the time lag since the jth reference to chunk i and d is the time decay parameter, typically set at 0.5.

- Retrieval from memory is computed by selecting the chunk with the highest activation value, after noise has been added. Analytically, the probability P_i of retrieving chunk i can be characterized by the Boltzmann (softmax) distribution as $P_i = e^{A_i/t}/\sum_j e^{A_j/t}$ where the sum is over all chunks j matching the retrieval request and the temperature t is a function of the noise parameter. This is equivalent to viewing the activation of a chunk as an estimate of the log odds of retrieval need (Anderson (1990)).

---

[1] https://github.com/dfmorrison/pyactup/

- The latency T_i of a chunk retrieval is inversely proportional to its activation as: $T_i = Fe^{-A_i}$ when F is a time scaling parameter.

Although attitudes have been modeled using ACT-R in prior work (Orr et al., 2021; Pirolli, 2016a,b; Pirolli et al., 2020), there exists no direct comparison to prominent models in the social psychology literature.

Data

We generated synthetic data for both studies in this article using two bit vectors as the basis for the synthetic data. The intent is for those vectors to represent two distinct attitudes competing in belief space. To generate the basis bit vectors, we used the following procedure: Take any random bit vector of length 16 with exactly eight bits with a state of 1 as the first pattern ζ^1. Then, generate another pattern ζ^2 from ζ^1 by flipping four of the 1 bits and four of the 0 bits. This procedure results in the two patterns ζ^1 and ζ^2 that are exactly the expected Hamming distance among all possible vectors in the configuration space of size 2^{16}. For ease of analysis, we fixed ζ^1 to 1111111100000000 and generated ζ^2 as 1111000011110000; these were our two basis bit vectors.

We constructed five sets of data, all using the same procedure. The basic unit of data was called an example, a single 16-bit vector. We first defined five frequency ratios, each mapping to one of the five sets of data: 50:50, 60:40, 70:30, 80:20, 90:10. The first term of each ratio referenced the number of examples of ζ^1 in the data set; the second term did the same for ζ^2. Each of the five sets of data also contained one example from the full configuration space of 2^{16} (that is 65,536 distinct examples define the configuration space). Thus, each of the five data sets contained a total of 65,636 examples, 100 of which were some ratio of ζ^1 and ζ^2.

The CAN model assumes that each node in a Hopfield network captures the endorsement or not of a belief that references an attitude object (e.g., 'has claws' is a belief about cats that is either endorsed or not). We use the same abstraction in our simulations and will call each bit in the bit vector an attitudinal belief.

Simulations

The two models (CAN and ACT-R) learned the data via a single pass through all examples in a data set (for both Study 1 and 2). The notion, in attitude research, is that each example is an abstraction of a social exposure to a set of beliefs (e.g., from an acquaintance or from mass media). We will call this the learning phase, which was identical in all conditions across Studies 1 and 2 (except for the distinct frequency distributions of each condition). We ran two separate studies.

Study 1: Frequency Effects in Free Recall. The objectives of Study 1 were to understand how each of the model types (CAN and ACT-R) represents differences in frequency of inputs and how this affects retrieval under free-recall. For each model type there were five conditions, one for each of the five data sets, which determined the data that the model

learned. Following learning, each model generated a non-cued retrieval probability for each of the 2^{16} bit vectors in the full configuration space. (See the section *Design* for computation of these probabilities.) Due to stochasticity in retrieval in ACT-R, we computed the set of retrieval probabilities for each model for each condition 30 times, the average of which was reported for the two basis patterns ζ^1 and ζ^2.

Study 2: Frequency Effects in Cued Recall. For Study 2, we used the same method as for Study 1 with one exception, cuing. In Study 2, we ran the full set of simulations used in Study 1 two separate times, each with a different cue. The first time used the more frequent basis pattern ζ^1 as the cue; the second time used the less frequent ζ^2.

The ACT-R Model: We defined all chunks to have one slot for each of the 16 attitudinal beliefs (16 bits in the bit pattern). Each slot had two valid values, 0 and 1. For the learning procedure, the model encoded all examples in its condition. The frequency of each chunk was reflected in the data so chunks were reinforced in proportion to their frequency by separate chunk encodings (i.e., each chunk was reinforced as many times as there were examples in the data). We used the functions `pyactup.learn()` to learn chunks and `pyactup.advance()` to advance time. All chunks were learned prior to advancing time and thus retrieval was not subject to time-dependent decay across chunks. For the simulation procedure we used the `pyactup.retrieve()` function. In Study 1, all retrievals were non-cued. For Study 2, each cue condition was realized by providing the cue of the full pattern of interest, either ζ^1 or ζ^2 e.g., `pyactup.retrieve({$\zeta^1$})`. All parameters of the cognitive architecture were left at their default values, i.e., the decay rate was 0.5, the activation noise was 0.25 and the retrieval threshold was 0.0.

The CAN Model: The Hopfield model was constructed by (i) mapping each of the bits x_i to a network node, (ii) generation of weights w_{ij} using Hebbian learning (Hertz et al., 1991), (iii) assigning a baseline parameter for each x_i as τ_i. Cuing (or not) was controlled by the set of τ_i. In Study 1, all τ_i were set to zero, to reflect no cuing free-recall. In Study 2, cuing was defined as providing the following mapping: $x_i = 1 \longmapsto \tau_i = 1$ if $x_i = 1$ was learned; else $x_i = 0 \longmapsto \tau_i = -1$; the latter condition provided a strong bias for $x_i = 0$.

Results

Study 1: Frequency Effects in Free Recall

The primary result in Study 1, shown in Figure 1, was the comparison between the ACT-R and CAN attitude models under no cuing conditions. Both models responded in a way that captured the frequency ratio between the basis patterns ζ^1 and ζ^2 (note: ζ^1 is more frequent). When the ratio was 50 : 50 the probability of recall was nearly equal between the two basis patterns for both models. As the ratio increased, for both models, the separation in probability of recall grew as a function of the size of the ratio between the two basis patterns. Two features distinguish the two models. First,

the CAN model had lower probabilities of retrieval overall. Second, also for the CAN model, the probability of retrieval for the less frequent basis pattern ζ^2 was very close to zero for any condition other than the 50 : 50 ratio. It is not clear whether these two features of the CAN model indicate a functional difference between it and the ACT-R model.

Figure 1: A comparison between the ACT-R (top panel) and CAN (bottom panel) attitude models in the probability of retrieval as a function of each of five conditions of the frequency ratio of the two basis patterns ζ^1 and ζ^2 (the former is 65280; the latter is 61680) in Study 1. No cue was given in this study. Note: ζ^1 is more frequent.

Figure 2 provides some insight into the way the models operate; it shows the results of a single simulation in the condition 50 : 50. Both models cleanly separated the two basis patterns ζ^1 and ζ^2 from the other patterns. For the CAN model, the point shown with the highest probability of retrieval captured the two basis patterns (this is occluded because of overlap). For ACT-R, one of the basis patterns was clearly favored, something that was due to the stochastic nature of activation noise in each chunk. Figure 3 shows results for the 80 : 20 condition. We see that with a high frequency ratio, both models showed strong separation of the most frequent basis pattern ζ^1. Comparing the two conditions (50 : 50 to 80 : 20) surfaces one potentially interesting difference between the two models in terms of their operation. For the CAN model, a larger frequency ratio between the two basis patterns significantly affected the range of the energy surface via reducing the minimum energy of the system (it deepened the attractor); the corresponding effect in terms of activation in ACT-R was much more muted.
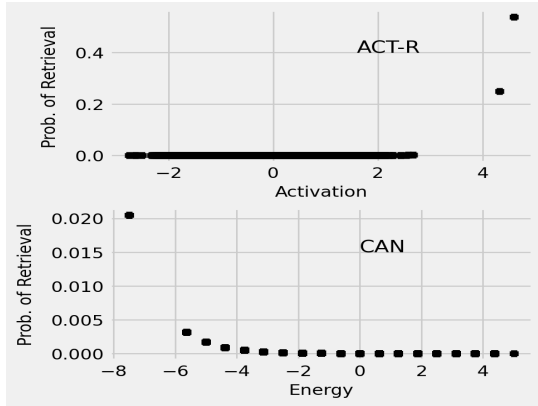
Figure 2: The relation between energy (CAN model) or activation (ACT-R model) (x-axis) and the probability of retrieval (y-axis) for each of the examples in the full configuration space (2^{16} examples). Each panel represents a simulation of the 50 : 50 condition. Note the different scales of the y-axis in each panel.
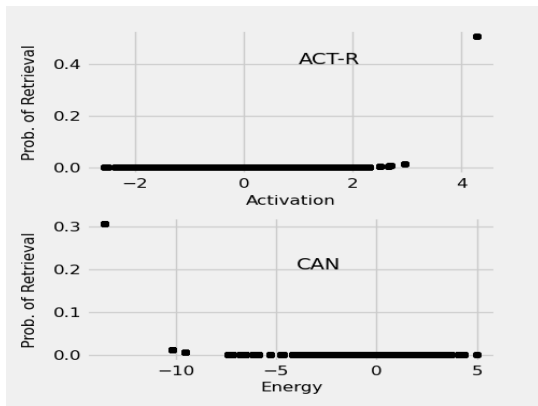
Figure 3: The relation between energy (CAN model) or activation (ACT-R model) (x-axis) and the probability of retrieval (y-axis) for each of the examples in the full configuration space (2^{16} examples). Each panel represents a simulation of the 80 : 20 condition. Note the different scales of the y-axis in each panel.

In summary, the first order comparison between the ACT-R and CAN attitude models showed functional similarity, to a first approximation, in terms of reflecting the frequencies of the learning environment (see Figure 1). Both models were good at separating the two basis patterns and their respective frequencies in terms of probability of retrieval (see Figures 2 and 3). The only notable difference, one for future study, was that the change in the energy space with an increased frequency ratio was much more significant for the CAN model than for the ACT-R model.

Study 2

The results for Study 2 were markedly different from Study 1. Figure 4 shows the set of simulations that cued the more frequent bit pattern ζ^1 (we will call these *Study 2a*). The high-level feature of these data is that both models operated well under cue in the sense that under all frequency ration conditions the cue was likely to be retrieved. This was to be expected because we cued the most frequent bit pattern. Further, the behavior of the ACT-R model was completely dependent on the cue; its behavior was the same for all five frequency ratios. In contrast, the CAN model exhibited a strong frequency effect across the frequency ratio spectrum. We will come back to this latter point shortly.

The set of simulations (*Study 2b*) that cued the less frequent bit pattern ζ^2 are shown in Figure 5. The comparison between ACT-R and CAN showed clear differences. As in Study 2a, the ACT-R model was completely driven by the cue and showed no effect across the frequency ratio conditions. In other words, the partial matching term overwhelmed the base-level activation, partly due to the large size of the fully-specified pattern (16 slots). However, for the CAN model we see see an interaction (of sorts) between the context of the cue and the frequency ratio of what was learned. For lower frequency ratios, the CAN model cued accurately but for higher frequency ratio conditions, the frequency factor drove the probability of retrieval. This, in fact, is the same effect we saw in Study 2a for the CAN model–the probabilities of retrieval decreased as the frequency ratio became smaller, conditions for which the learning context was against, in a relative sense, the more frequent bit pattern ζ^1.

In summary, in both Study 2a and 2b, we see a strong frequency effects for the CAN model and not for the ACT-R model under cuing conditions.
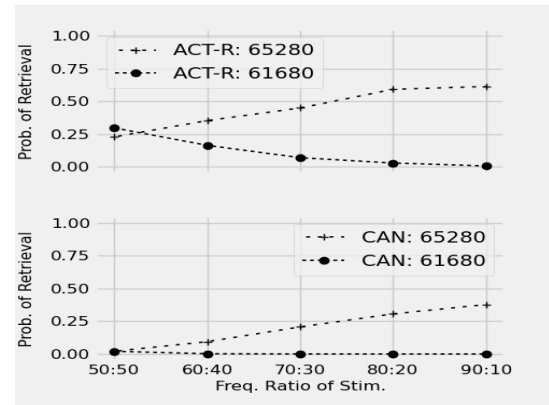

Figure 4: A comparison between the ACT-R (top panel) and CAN (bottom panel) attitude models in the probability of retrieval as a function of each of five conditions of the frequency ratio of the two basis patterns ζ^1 and ζ^2 (the former is 65280; the latter is 61680) in Study 2a. The cue was the more frequent pattern ζ^1.
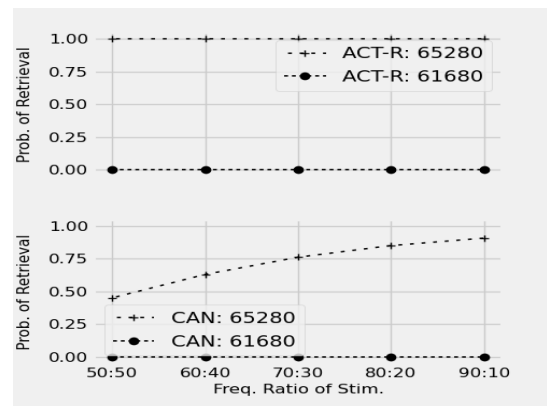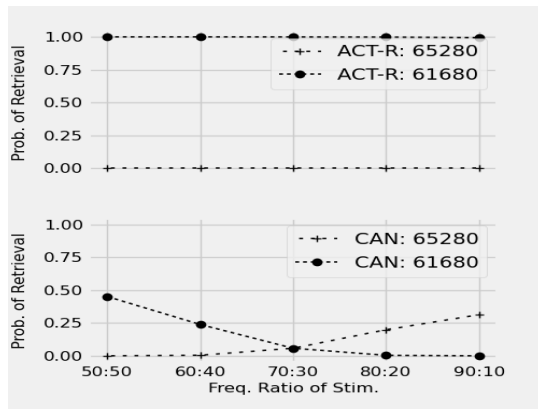
Figure 5: A comparison between the ACT-R (top panel) and CAN (bottom panel) attitude models in the probability of retrieval as a function of each of five conditions of the frequency ratio of the two basis patterns ζ^1 and ζ^2 (the former is 65280; the latter is 61680) in Study 2b. The cue was the less frequent pattern ζ^2.

Conclusions

In conclusion, the differences in the retrieval behavior of the ACT-R and CAN attitude models were greater than the similarities. Cued recall, a more realistic conceptualization of the human attitude problem, showed marked differences between the two models. The ACT-R attitude model was driven by the cue; the CAN model was driven by both cue and learning frequency, sometimes to the extent that the cue was effectively ignored. Although ignoring cues can be adaptive in some tasks, we do not see the value in the context of attitudes unless other social processes or motives were modeled in conjunction.

To what extent does this stand as an indictment of the CAN attitude model? On one hand, the declarative memory model in ACT-R could be seen to serve as a kind of validation comparison: it represents human memory in a way that is not justifiable for the CAN model. In the CAN model's defense, we note that the CAN model was not developed in the context of memory models. The CAN model was an outgrowth of what is called the psychological networks approach, an approach for using graph structure as an alternative measurement approach for psychological survey or clinical data.

We see our work presented here as highly provisional, a useful first step in reconciling learning to the structural approach to attitudes. Future work should study the following issues: (i) the degree of learning in the Hopfield network would impact the results, yet it is not clear to what extent or precisely how, (ii) formal mathematical analysis and comparison of learning and retrieval in both the ACT-R and CAN models, (iii) evaluating the impact of different representations in the ACT-R model, e.g., by representing each belief as a separate chunk, (iv) whether the results generalize to partial cueing of a subset of the full belief set, and (v) the impact of factors such as recency if a real time learning schedule is

used.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036.

Corneille, O., & Stahl, C. (2019). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*, *23*(2), 161–189.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. (2016). Toward a formalized account of attitudes: The causal attitude network (can) model. *Psychological review*, *123*(1), 2.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2018, Oct). The attitudinal entropy (ae) framework as a general theory of individual attitudes. *Psychological Inquiry*, *29*(4), 175–193. doi: 10.1080/1047840X.2018.1537246

Dalege, J., & van der Maas, H. L. J. (2020, Nov). Accurate by being noisy: A formal network model of implicit measures of attitudes. *Social Cognition*, *38*(Supplement), s26–s41. doi: 10.1521/soco.2020.38.supp.s26

Eagly, A., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace Jovanovich.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Addison-Wesley.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Orr, M., Stocco, A., Lebiere, C., & Morrison, D. (2021). Attitudinal polarization on social networks: A cognitive architecture perspective. In *Proceedings of the 19th international conference on cognitive modelling*.

Pirolli, P. (2016a). A computational cognitive model of self-efficacy and daily adherence in mhealth. *Translational behavioral medicine*, *6*(4), 496–508.

Pirolli, P. (2016b). From good intentions to healthy habits: Towards integrated computational models of goal striving and habit formation. In *2016 38th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 181–185).

Pirolli, P., Bhatia, A., Mitsopoulos, K., Lebiere, C., & Orr, M. (2020). Cognitive modeling for computational epidemiology. In *2020 international conference on social computing, behavioral-cultural modeling & prediction and behavior representation in modeling and simulation (sbp-brims 2020)*.

Predicting Learning and Retention in a Complex Task

David Peebles (d.peebles@hud.ac.uk)
Department of Psychology, University of Huddersfield,
Queensgate, Huddersfield, HD1 3DH, UK

Abstract

This paper reports an experiment investigating learning and retention in a complex task over multiple sessions across an extended period of time. The primary aim of the experiment is to evaluate the *Predictive Performance Equation* (PPE: Jastrzembski & Gluck, 2009) a model of learning and forgetting that predicts retention based on past performance. The second aim is to test a taxonomy for knowledge, skills and attitudes and a competence retention analysis technique developed to improve competence retention in military training (Cahillane, Launchbury, MacLean, & Webb, 2013). Participants were trained over 16 weeks on the Multi-Attribute Task Battery (MATB: Comstock Jr & Arnegard, 1992), a computer-based task analogous to piloting an aircraft. The study reveals significant variation in learning profiles for the MATB subtasks and demonstrates the PPE's ability to make accurate predictions of human performance over intervals ranging from 27 to 111 days.

Keywords: MATB; Predictive Performance Equation

Competence retention and training

Many military personnel are required to maintain high levels of task knowledge and skill performance and so are subjected to regimes of regular testing and refresher training to combat the effects of skill fade. The schedule of retraining is typically not determined on an individual basis but is standardised (e.g., calendar-based) and the acceptable threshold criterion is either a general numerical measure such as the number of training hours completed or a qualitative "pass/fail" score. However, because there are substantial differences in people's ability to learn and retain information, it may be the case that two individuals with the same training schedules perform (possibly safety or mission critical) tasks at very different levels of effectiveness.

To complicate matters, there is strong evidence from the psychological literature that knowledge and different types of skills decay at different rates (e.g., Wisher, Sabol, & Ellis, 1999; Stothard & Nicholson, 2001). Together, these two factors suggest that a more efficient and productive approach to training and skill maintenance would be to derive personalised training schedules through detailed analysis of the knowledge, skills and attitudes involved in the task and from each individual's learning and retention profile.

This paper describes an experimental study that aims to investigate and integrate two approaches to the understanding and improvement of competence retention and the personalisation of learning for Defence. The first involves the application of a model of learning and retention called the *Predictive Performance Equation* (PPE)—to create personalised training schedules based on predicted memory retrieval failure (Jastrzembski & Gluck, 2009). The second approach relates to research conducted by the UK Defence Science and Technology Laboratory (Dstl) to develop a set of principles for improving competence retention in military training, together with a *competence retention analysis* (CRA) technique to support competence retention through training (Cahillane et al., 2013). The paper will proceed by first describing the two strands of research, then outlining the details of the experiment, and finally discussing some of the key results, implications and limitations of the study.

The Predictive Performance Equation

The acquisition and retention of knowledge are influenced by three primary factors: the amount of practice (the frequency effect), the amount of time elapsed since the last practice session (the recency effect), and the temporal distribution of practice (the spacing effect). The spacing effect is less intuitive than the others but is a ubiquitous occurrence in learning in which practice sessions which are more widely distributed over time result in better retention compared to identical training sessions scheduled closer together. The beneficial effect of increasing the study interval works only up to a certain point; intervals beyond a certain threshold diminish final retention (Benjamin & Tullis, 2010), but it has been argued that informed use of the spacing effect can have significant positive implications for education and training (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012).

The PPE characterises the combined effects of recency, frequency and spacing on retention and subsequent task performance (Jastrzembski & Gluck, 2009; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018). When calibrated to individual task performance data gathered over a series of sessions, the PPE is able to account not only for existing performance data but is also able to make precise, quantitative predictions of an individual's performance at later points in time, sometimes many months into the future (Jastrzembski et al., 2017). It does so by calculating the expected stability of knowledge and skills on the basis of the previous training history and using this measure to predict the expected retention of knowledge and skills across periods of non-use or further practice (Jastrzembski, Portrey, Schreiber, & Gluck, 2013). The PPE is able to provide an accurate estimate of the time when performance has declined to such an extent that refresher training is required.

A key premise of the PPE is that learning new information creates traces in Long-Term Memory (LTM) and that each trace has a degree of *activation*, which determines the probability that it can be subsequently retrieved and the speed with which that retrieval will be accomplished. The mechanism of

activation to explain the effect of elapsed time and practice on task performance. According to this account, LTM traces vary in their base level of activation (often referred to as the "strength" of the trace) depending on how frequently or recently they have been used and a trace's strength determines its general availability. The strength of a trace changes gradually; it decays over time but can be progressively increased by repeated practice. The availability of a memory trace is what affects performance.

The PPE equation (Equation 1) predicts the activation of a memory trace, which is subsequently converted to a performance prediction.

$$M_n = (N + a)^c \cdot T^{-d} \quad (1)$$

The PPE assumes that performance increases as a power function (with learning rate c) of the number of practice episodes N, decreases as a power function (with decay rate d) of elapsed time (in seconds) since the episodes occurred T, and that the effects of practice and elapsed time are multiplicative in nature (although, as described above, this effect is on the activation, M of a memory trace, n rather than on performance directly). Finally the a parameter represents an individual's prior experience in the task, adding to the number of practice episodes to increase ease of activation.

In order to incorporate the spacing effect into the equation, a conception of time is introduced in which T is computed as the sum of the time (t_i represents the time of encounter i) since each of the previous study or practice events, each weighted, w_i so that the most recent events are given extra prominence—the older an encounter the smaller the contribution of that encounter to the total time.

$$T = \sum_{i=1}^{n-1} w_i \cdot t_i \quad (2)$$

Weight is an exponential decay function of time according to Equation 3. In this equation, shorter time distances (i.e., more recent encounters with the material) are weighted more heavily, with the x parameter determining the degree of prominence given to shorter time delays. The x parameter is typically set to 0.6 as this provides a good fit to data in many studies.

$$w_i = \frac{t_i^{-x}}{\sum_{j=1}^{n} t_j^{-x}} \quad (3)$$

Equation 4 determines the rate of decay, d for memory traces and is defined to capture the spacing effect in which longer delays between practice episodes result in a reduction in decay rate—and as a consequence produce more stable knowledge (Jastrzembski & Gluck, 2009; Walsh, Gluck, Gunzelmann, Jastrzembski, Krusmark, Myung, et al., 2018).

As the interval between study and test increases, the lower decay rate associated with spaced versus massed repetitions enhances retention. To capture this effect, the function incorporates the history of lags (i.e., time differences) between successive events and is a linear function of the average of one over the sum of the natural logarithm of the lags. In this equation, b and m are parameters that determine the decay intercept and slope of the function and correspond to an individual's overall level of forgetting and their susceptibility to the spacing effect respectively. When lags are long, the value inside the brackets approaches zero, reducing decay to the asymptotic value determined by the b parameter. In contrast, when the lags are short the value inside the brackets approaches one, increasing decay.

$$d_n = b + m \cdot \left(\frac{1}{n-1} \cdot \sum_{j=1}^{n-1} \frac{1}{\log(lag_j + e)} \right) \quad (4)$$

Finally, the level of activation, M_n, computed in Equation 1 is transformed into a continuous response value that represents performance, P_n according to Equation 5. Performance is a logistic (sigmoid) function of activation with range $[0, 1]$ where the τ parameter determines the sigmoid's midpoint and the s parameter determines the logistic growth rate (i.e., the steepness of the curve).

$$P_n = \frac{1}{1 + \exp\left(\frac{\tau - M_n}{s}\right)} \quad (5)$$

The PPE has been tested in several studies to determine its ability to predict skill fade and when individuals need to return for retraining on critical tasks (Jastrzembski, Gluck, & Rodgers, 2009; Jastrzembski et al., 2013; Gluck et al., 2019; Jastrzembski et al., 2017) and the results so far indicate that the PPE is able to track and predict performance accurately at the individual learner level over timescales ranging from seconds to months.

Competence Retention Analysis

Competence retention analysis (Cahillane et al., 2013) is a novel approach developed by the UK Ministry of Defence (MoD) aimed at formulating a set of generic principles and guidance for the optimisation of competence retention in military training. To achieve this, a new classification of the knowledge, skills and attitudes (KSA) was developed that was consistent with the current psychological literature on mechanisms underlying competence retention and their differential rates of decay.

The primary aim of the CRA is to be a framework grounded in psychological evidence that can provide generic advice and guidance for training designers. Once tasks have been analysed in terms of their cognitive components, designers can consult the CRA to determine the likely retention profiles for the individual components and the task as a whole and then plan refresher training schedules accordingly.

The CRA is based on a three-level categorisation of retention, defined on a criterion value of 50% competence after a given period of time since the last training session. According to this classification, a "high" level of retention is greater than 50% competent after 12 months non-practice, a "moderate" level is 50% competent after 5 months non-practice, and

a "low" level of retention is 50% competent after two months non-practice.

In addition, the relationship between psychological components and retention categories can be moderated by the frequency with which they are applied when performing a given task. The CRA defines three frequency levels: "very frequent" (more than once every two months), "moderately frequent" (between once every two months and once every five months), and "infrequent" (once in a period greater than five months). The resulting taxonomy consists of a knowledge domain and four types of skill:

- **Explicit knowledge.** Knowledge required to conduct a task, such as facts, concepts and theories. Retention: High, Frequency: Infrequent.

- **Continuous psychomotor skills.** Tasks requiring the ability to perform well-trained and practiced motor actions that do not have distinct beginnings or endings (e.g., driving, flying an aircraft and target tracking). Retention: High, Frequency: High/Moderate.

- **Discrete psychomotor skills.** Physical tasks with discrete beginnings and endings that rely on both procedural and perceptual motor skills (e.g., disassembling a weapon or other weapon handling tasks). Retention: High, Frequency: High/Moderate.

- **Procedural skills.** Tasks requiring working memory to remember a sequence of steps and their order nature (e.g., using a Battlefield Information Management System (BIMS) to create map overlays (Cahillane & Morin, 2012)). Retention: Low, Frequency: High/Moderate/Infrequent.

- **Decision making skills.** Tasks involving the application of cognitive processes such as, judgement, problem solving and analysis in order for an individual to arrive at a decision (e.g., troubleshooting faulty equipment). Retention: Moderate, Frequency: Infrequent.

Experiment

To reiterate, the experiment has two aims. The first is to determine whether the retention profiles of individuals engaged in a complex task can be captured by the PPE to allow accurate prediction of future performance. The second is to investigate the learning and retention profiles of tasks involving the different psychological domains identified by the CRA.

To achieve both aims, the task selected for the experiment was the Air Force Multi-Attribute Task Battery (AF-MATB; Comstock Jr & Arnegard, 1992; Miller, Schmidt, Estepp, Bowers, & Davis, 2014). MATB is a computer-based interactive multitasking environment consisting of a set of four subtasks designed to be analogous to those performed during aircraft piloting. It has been widely used to study the effects of various factors (e.g., automation, priorities, instructions, task difficulty, etc.) on a range of behavioural measures, including multitasking, attention management, vigilance, decision

making, ocular behaviour, prospective memory and subjective mental workload. Crucially for this study, MATB is relevant to Defence and consists of multiple components involving different skill domains where performance can be quantitatively measured. In addition, it has been demonstrated that people learn and improve over time during the task (e.g., Fairclough, Venables, & Tattersall, 2005; Kee et al., 2019).
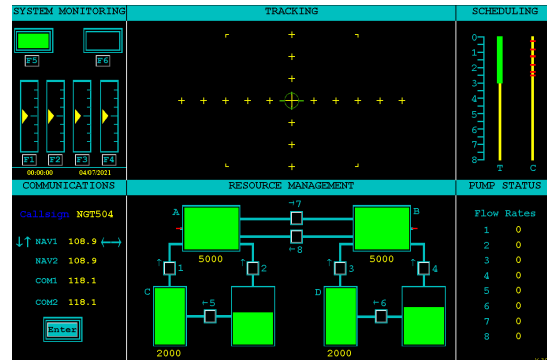
Figure 1: The AF-MATB task interface.

A detailed description of the AF-MATB can be found in Miller et al. (2014) but to summarise, the display consists of four task windows and two information windows (shown in Figure 1). The four tasks and their performance measures are:

- **System monitoring (SYSMON).** Participants monitor gauges and warning lights and must respond to changes by pressing an appropriate key within a time interval. Performance is measured as the proportion of correct responses.

- **Tracking (TRACK).** Participants must use a joystick to keep a randomly moving cursor inside a target area. Performance is measured as the root mean square deviation (RMSD) distance between the central crosshair and target.

- **Communication (COMM).** Participants must respond to specific auditory messages by adjusting radio and frequency values based on the message. Performance is measured as the proportion of correct adjustments.

- **Resource management (RESMAN).** Participants must maintain the fuel tank levels within target ranges by turning on or off a set of pumps. Performance is measured as the root mean square deviation (RMSD) between actual and target fuel levels.

For the purposes of this study, three of the subtasks were associated with CRA skill domains: the TRACK task with the continuous psychomotor (high retention) domain, the RESMAN task with the decision making (moderate retention) domain, and the COMM task with the procedural (low retention) domain. To the extent that these subtasks require the use of a particular CRA cognitive domain, it is expected that their retention profiles will differ. Specifically, the TRACK task should be retained better than the RESMAN task which in turn should be retained better than the COMM task.

(a) COMM (CM-P) (b) RESMAN (RM-P) (c) TRACK (TR-P)

(d) SYSMON Gauge (SG-P) (e) SYSMON Light (SL-P) (f) Performance ranges

Figure 2: 2a–2e: Learning profiles for the three training schedule conditions over the four training sessions. 2f: Range of performance (P) and task completion times (RT) values for each subtask. Error bars indicate standard deviation.

Participants and materials
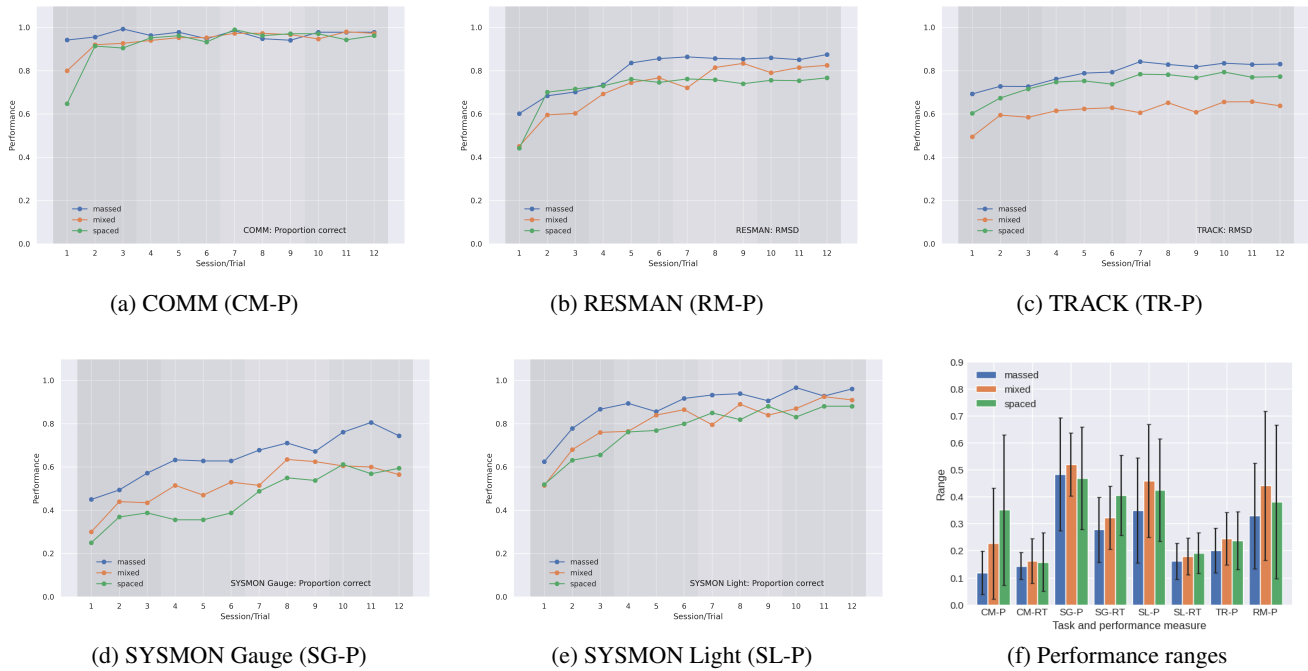
Participants were 27 staff, faculty and students from the University of Huddersfield. All participants were 18 years old or over, had normal or corrected to normal eyesight, and were paid £10 per session. The experiment was conducted on PCs running Microsoft Windows 10 with 24-inch displays at 1080p resolution. Participants interacted through the computer keyboard and a Logitech G Extreme 3D PRO joystick.

Design and procedure

The experiment lasted 16 weeks and consisted of four training sessions followed by two assessment sessions[1]. There were three training schedule conditions. Participants in the "massed" condition (9 in total) attended once a day for 4 consecutive days in Week 4. Participants in the "spaced" condition (8 in total) attended once a week (on the same day) for 4 consecutive weeks, while participants in the "mixed" condition (10 in total) attended twice a week for 2 alternate weeks.

For all conditions, the fifth and sixth testing sessions were approximately 42 days (Week 10) and 84 days (Week 16) after the last training session. The aim of creating different scheduling conditions was to provide variation in training spacing as this is a key determinant of the PPE model's predictions. Because the spacing effect is a very well-established result however (e.g., Latimier, Peyre, & Ramus, 2021), this

experiment was not designed to include statistical analysis of any effects of spacing on human performance.

MATB event schedules were created to define three 10-minute trials which were different in terms of their event scheduling but equal in difficulty (i.e., number of events and the degree of multitasking required to process them). The trials were designed to be challenging to enable continued performance improvements over the sessions and minimise the likelihood of performance reaching ceiling. A previous MATB study of learning found that sustained learning was only observed in trials where the task demand was high (Fairclough et al., 2005).

Before the first session, participants were introduced to the MATB task via a video which explained the four subtasks and how to interact with the software. Training and assessment sessions consisted of three 10-minute trials (in random order) separated by rest intervals of up to 5 minutes. After each trial, the experimenter would restart the software and ensure that participants had a break and were ready to continue.

Results

Space limitations preclude a full account of the analyses here but further details and data are provided on the study's OSF web page. The sections below will first describe the results of applying the PPE to predicting individual performance in the two test sessions and then report the learning profiles for the MATB subtasks over the four learning sessions.

PPE predictions of human performance While it is possible to apply the PPE to predict performance on the individ-

---

[1]The study was preregistered with the Open Science Framework (osf.io/uc4fy) and was approved by the Research Ethics Committees of the Ministry of Defence and the University of Huddersfield School of Human and Health Sciences

Table 1: Comparison between human performance and model predictions, sessions 5 and 6

		Session 5						Session 6					
P	Schedule	Human	Model	R^2	RMSE	Days	4-5 Diff	Human	Model	R^2	RMSE	Days	5-6 Diff
1	Mixed	0.654	0.644	0.931	0.024	42	-0.031	0.627	0.618	0.881	0.031	42	0.016
2	Mixed	0.620	0.581	0.887	0.032	42	0.003	0.623	0.580	0.857	0.036	42	-0.022
3	Mixed	0.645	0.636	0.155	0.034	43	-0.059	0.583	0.601	0.010	0.052	48	-0.071
4	Mixed	0.568	0.589	0.518	0.026	42	0.004	0.583	0.589	0.428	0.028	42	-0.011
5	Mixed	0.613	0.672	0.959	0.019	42	0.117	0.624	0.662	0.938	0.023	43	0.100
6	Mixed	0.856	0.846	0.949	0.008	63	-0.036	0.855	0.844	0.895	0.012	28	0.000
7	Spaced	0.553	0.610	0.920	0.025	42	0.084	0.545	0.585	0.842	0.030	42	0.015
8	Spaced	0.495	0.620	0.923	0.017	42	0.080	0.543	0.553	0.695	0.031	42	0.018
9	Mixed	0.673	0.735	0.932	0.019	42	0.035	0.694	0.702	0.822	0.028	41	0.007
10	Mixed	0.702	0.695	0.949	0.021	41	-0.037	0.693	0.688	0.934	0.023	43	0.024
11	Mixed	0.754	0.751	0.863	0.016	41	-0.026	0.760	0.734	0.720	0.021	43	-0.051
12	Spaced	0.729	0.731	0.918	0.025	56	-0.002	0.746	0.712	0.847	0.034	27	-0.014
13	Spaced	0.772	0.809	0.844	0.016	42	0.028	0.781	0.788	0.724	0.020	42	0.028
14	Mixed	0.719	0.728	0.911	0.020	41	0.071	0.778	0.721	0.887	0.022	43	-0.031
15	Spaced	0.761	0.769	0.970	0.014	42	0.026	0.757	0.761	0.943	0.019	42	0.030
16	Spaced	0.797	0.777	0.954	0.014	42	0.007	0.749	0.769	0.913	0.019	42	0.033
17	Spaced	0.820	0.746	0.512	0.026	42	-0.111	0.786	0.752	0.431	0.031	42	-0.022
18	Spaced	0.552	0.576	0.855	0.027	42	0.003	0.551	0.568	0.823	0.030	42	0.036
19	Massed	0.756	0.732	0.877	0.021	50	-0.043	0.708	0.728	0.857	0.022	41	0.041
20	Massed	0.786	0.784	0.623	0.020	42	0.026	0.790	0.787	0.648	0.020	42	0.026
21	Massed	0.679	0.775	0.896	0.025	42	0.087	0.734	0.711	0.740	0.036	42	-0.036
22	Massed	0.783	0.803	0.946	0.017	49	0.008	0.778	0.775	0.916	0.020	35	0.005
23	Massed	0.605	0.666	0.899	0.027	47	0.044	0.615	0.616	0.846	0.032	37	0.012
24	Massed	0.784	0.803	0.738	0.011	42	0.007	0.809	0.794	0.723	0.011	42	0.000
25	Massed	0.748	0.769	0.834	0.026	42	0.041	0.729	0.750	0.817	0.026	42	0.063
26	Massed	0.785	0.811	0.693	0.015	42	0.027	0.791	0.806	0.664	0.016	35	0.051
27	Massed	0.771	0.787	0.935	0.015	35	0.004	0.756	0.743	0.838	0.023	111	0.002
	Mean	0.703	0.720	0.829	0.021	43.7	0.013	0.703	0.701	0.764	0.026	43.1	0.009
	StDev	0.095	0.080	0.186	0.006	5.4	0.050	0.091	0.084	0.202	0.009	14.3	0.036

ual MATB subtasks, for this study the subtask measures were transformed onto a common scale and then averaged for each participant to create a single, global MATB score.

The PPE's predictions were tested on sessions 5 and 6, approximately 43 and 86 days respectively after a participant's fourth training session. For each test, the model was fitted to the individual's performance data from the previous sessions by adjusting five free parameters: b and m, representing the intercept and slope of the decay function (Equation 4) respectively, τ and s which determine the intercept and slope of the activation transformation function (Equation 5) respectively, and a representing an individual's prior experience (Equation 1). The fitted model was then used to predict performance at the date and time of the first trial of the test session.

Table 1 displays the results of the modelling, showing participants' performance, model predictions, and the difference in participants' performance from the last trial of the previous session and the first trial of the current session. Participants' performance varies widely in both sessions (e.g., compare participants 6 and 8) but there was typically little change in performance between sessions, despite a mean interval of approximately 43 days, indicating that, in general, retention remained stable. With a few notable exceptions (e.g., participants 3 and 17 who showed little decay in performance or, somewhat counterintuitively, performance improvements, after time delays), the PPE was able to provide a close fit to the data and make accurate predictions beyond the training set.

Subtask learning profiles Figure 2 shows the learning profiles for the MATB subtasks over the 12 trials of the four

learning sessions. All performance measures are scaled to the range [0,1] to allow comparison. Figure 2f depicts the range of performance scores and task completion times produced by each schedule condition for the different subtasks. For example, the low performance range for the TRACK task reflects the relatively shallow learning curves, in contrast to the much greater changes found in the SYSMON gauge task.

Although the relatively small number of participants limits comparison of the schedule conditions, interesting features can seen in the individual subtask data for all three. First, there was a general level of consistency in performance between the three training schedules, not only in the ranges of values produced but also in the performance profiles across the training phase. Performance differences were also evident in the four subtasks. For example, participants in all three conditions quickly achieved and maintained very high levels of accuracy in the COMM task, whereas in the RESMAN task, performance increased more gradually by approximately 30% to 40% during the course of training.

These differences are likely due to the nature of the interactions required. For example, the time constraints of the COMM task demand immediate attention and a rapid sequence of actions to encode, retain, and then enter information into the system, a task that participants cannot complete much faster than 4.5 seconds. Performance improvements in the other subtasks are likely to be due to, amongst other things, the refinement of local and global strategies, for example revising priorities when balancing different resources in the RESMAN task and more efficiently allocating attention

when managing competing demands from subtasks.

Discussion

This experiment has generated a rich dataset of individual learning and forgetting in a complex task involving multiple sub-tasks which is yet to be fully analysed. The main analysis reported here however provides additional support for the PPE by demonstrating its ability to predict performance accurately over retention intervals ranging from 27 to 111 days. While the limited number of participants precludes rigorous statistical analysis of the training schedule conditions, the different subtask learning profiles do provide valuable initial pointers for further investigation. While the pattern of differences in learning are not consistent with the classification provided by the CRA, additional analysis of the data from sessions 5 and 6, combined with a detailed task analysis, may provide further insight into differences in retention over longer intervals.

Acknowledgments

References

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*(3), 228–247.

Cahillane, M. A., Launchbury, C., MacLean, P., & Webb, S. (2013). *Competence retention* (Tech. Rep. No. DHC-STC_12_T_T2_001_1.1/005 V5.0). Defence Science and Technology Laboratory.

Cahillane, M. A., & Morin, C. (2012). Skills retention in a complex battlefield management system: A pilot study. *Journal of Battlefield Technology*, *15*(1), 65.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*(3), 369–378.

Comstock Jr, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research* (Tech. Rep. No. NASA-TM-104174). National Aeronautics and Space Administration.

Fairclough, S. H., Venables, L., & Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. *International Journal of Psychophysiology*, *56*(2), 171–184.

Gluck, K. A., Collins, M., Krusmark, M., Sense, F., Maaß, S., & van Rijn, H. (2019). Predicting performance in cardiopulmonary resuscitation. In *Proceedings of the 17th international conference on cognitive modelling. ICCM 2019.* Montreal, Canada.

Jastrzembski, T. S., & Gluck, K. A. (2009). A formal comparison of model variants for performance prediction. In A. Howes, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 9th international conference on cognitive modeling.* Manchester, UK.

Jastrzembski, T. S., Gluck, K. A., & Rodgers, S. M. (2009). The predictive performance optimizer: An adaptive analysis cognitive tool for performance prediction. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 53, pp. 1642–1646).

Jastrzembski, T. S., Portrey, A. M., Schreiber, B. T., & Gluck, K. A. (2013). Improving military readiness: Evaluation and prediction of performance to optimize training effectiveness. In W. Arthur, Jr., E. A. Day, W. Bennett, Jr., & A. M. Portrey (Eds.), *Individual and team skill decay* (pp. 177–199). Routledge.

Jastrzembski, T. S., Walsh, M. M., Krusmark, M., Kardong-Edgren, S., Oermann, M., Dufour, K., . . . Stefanidis, D. (2017). Personalizing training to acquire and sustain competence through use of a cognitive model. In D. Schmorrow & C. Fidopiastis (Eds.), *Augmented cognition. Enhancing cognition and behavior in complex human environments. AC 2017* (Vol. 10285). Springer.

Kee, T., Weiyan, C., Blasiak, A., Wang, P., Chong, J. K., Chen, J., . . . Asplund, C. L. (2019). Harnessing CU-RATE.AI as a digital therapeutics platform by identifying N-of-1 learning trajectory profiles. *Advanced Therapeutics*, *2*(9), 1900023.

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, *33*, 959–987.

Miller, W. D., Schmidt, K. D., Estepp, J. R., Bowers, M., & Davis, I. (2014). *An updated version of the US Air Force multi-attribute task battery (AF-MATB)* (Tech. Rep. No. AFRL-RH-WP-SR-2014-0001). Air Force Research Laboratory.

Stothard, C., & Nicholson, R. (2001). *Skill acquisition and retention in training: DSTO support to the army ammunition study* (Tech. Rep. No. DSTO-CR-0218). Electronics and Surveillance Research Laboratory, Defence Science & Technology Organisation.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., . . . Zhou, R. (2018). Mechanisms underlying the spacing effect in learning: A comparison of three computational models. *Journal of Experimental Psychology: General*, *147*(9), 1325–1348.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T. S., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive Science*, *42*, 644–691.

Wisher, R. A., Sabol, M. A., & Ellis, J. A. (1999). *Staying sharp: Retention of military knowledge and skills* (Tech. Rep. No. 39). US Army Research Institute for the Behavioral and Social Sciences.

How to Match Cognitive Model Predictions with EEG data

Kai Preuss (preuss at tu-berlin.de)
Cognitive Modelling in Dynamic Human-Machine Systems, Technische Universität Berlin, Marchstraße 23, 12051 Berlin

Christopher Hilton (c.hilton at tu-berlin.de)
Klaus Gramann (klaus.gramann at tu-berlin.de)
Biological Psychology and Neuroergonomics, Technische Universität Berlin, Fasanenstraße 1, 10623 Berlin

Nele Russwinkel (nele.russwinkel at uni-luebeck.de)
Institute of Information Systems, Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck

Abstract

Reliably identifying relevant brain areas implicated by the simulated activity from cognitive models is still an unsolved problem for cognitive modeling, particularly when matching model output with human electroencephalography (EEG) data. We propose a new method involving post-processing of ACT-R module activity and clustered EEG component activity, and performing generalized least squares (GLS) analysis to find matching patterns between predicted and observed data, thereby inferring neural substrates of distinct cognitive processes. This approach holds several advantages over other methods by controlling for autocorrelation and unequal variances. To exemplify its application, we used a cognitive model and EEG data from a mental spatial transformation study to show how this method finds areas involved in representational and transformational spatial processing. Parietal areas involved with spatial activity were identified, in line with prior studies on spatial cognition. In addition, previously established associations between ACT-R and brain areas were confirmed. Finally, we discuss limitations and possibilities of the approach.

Keywords: Electroencephalography; cognitive modeling; independent component analysis; generalized least squares; mental spatial transformation

Introduction

Simulating the activation strength of specific, functional areas of the human brain constitutes a complex problem. Biologically plausible cognitive architectures, such as ACT-R (Anderson et al., 2004), facilitate the creation of cognitive models that can produce differentiated activity during problem solving, which can in turn be related to functional brain areas. Establishing associations between simulations of cognitive processing and neural activity opens up a wide spectrum of possibilities, including further refinement of cognitive models to improve predictability and accuracy, assigning functional interpretability to neural substrates, and disambiguation of higher-order cognitive functions.

In the case of ACT-R, cognition is simulated by so-called *modules*, which independently process symbolic information and interact through a procedural production choice system. Module activity produced in ACT-R therefore lends itself to comparisons with neuroimaging data. These comparisons are usually performed on pre-selected neural substrates from fMRI data (e.g., Anderson et al., 2008; Borst & Anderson, 2017). Matching modeled predictions with recorded EEG data in a data-driven manner is rarely done, and no established methodology seems to be agreed upon. Exploratory search for EEG correlates of model output is a compelling possibility, as complex tasks may rely on higher-order cognition that might not be reproduced sufficiently by established cognitive processes only, and may require high temporal resolution to allow for its analysis. High task complexity can impede the application of prior knowledge or hypotheses about involved neural substrates, at which point an EEG-based, data-driven approach could deliver insight.

Only a few methods have so far been proposed for associating cognitive model output to areas of EEG activity. Canonical correlation was used by van Vugt (2012) to relate the time series of four ACT-R modules during an attentional blink task to EEG activity in six different frequency bands, which produced topography maps for model activity correlates. An early outline by Prezenski and Russwinkel (2016) suggested the use of cross-correlation for model-human data comparison. Still, these methods are not perfect: Anderson et al. (2008) outlined the difficulty of model-brain comparisons for complex cognition tasks, with difficulty of model-fit assessment, temporal variability between datasets and neglect of autoregression inherent in time series highlighted as the main problems. The authors suggest to mitigate these issues by minimizing the squared deviation between observed and predicted data, using event-locked data, and controlling for autocorrelation of predictions errors, respectively.

Generalized least squares regression (Lawson & Hanson, 1995) has proven to be a promising variant of linear models for brain activity analysis (e.g., Katanoda et al., 2002; Sato et al., 2006). It allows for the definition of data-specific correlation structures, which control for autocorrelation of the predictors, and variance structures, which control for unequal (condition-specific) variance. Compared to ordinary least squares regression, this increases its robustness with temporally or spatially correlated data. As such, it is especially suited for gauging the effect of multiple time series on one another.

Using GLS regression, patterns in module activity produced by cognitive models can be matched to patterns in EEG activity by fitting a model where the latter is predicted by the first. Significant predictivity by trial- or group-level EEG signals of module activity can then suggests a functional link between module and neural substrate. In turn, this reveals module-brain correlates that are not reliant on prior assignment of functionality to specific brain regions.
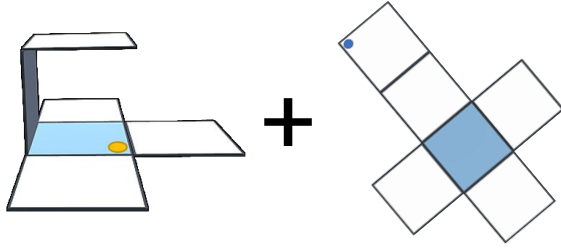
Figure 1: Example for the combined mental rotation and folding task.

In this paper, we present an approach of finding correlates between cognitive modeling and EEG data using GLS regression. We demonstrate the feasibility of this method by applying it to a mental spatial transformation study: in a combined mental rotation and folding task (cMRF) based on the mental rotation (Shepard & Metzler, 1971) and mental folding (Shepard & Feng, 1972) paradigms, participants gauged if a marker on two simultaneously presented figures each would align after spatial transformation was applied. During the experiment, EEG activity was recorded. In addition, a cognitive model for the task was created in ACT-R, and activity predictions of individual modules during trials were generated. Our aim was the characterization of spatial processes, the identification of their neural substrates, and the differentiation of these processes by storage- and manipulation-related cognitive loads. We will show how to apply the methodology, how it can be further adapted, what caveats remain, and provide an example for how this approach succeeded in its goal.

Methodology

Experiment

37 participants took part in the experiment, solving 610 trials in 5 blocks, in a pseudo-randomized order. After a fixation cross, the reference figure was presented, which was either a two-dimensional cube net or a partially folded three-dimensional structure, depending on the experiment condition. The base square was marked in blue and contained a yellow marker in one of the square's corners. One second later, the target figure was presented alongside the reference stimulus. The target was a two-dimensional cube net in all conditions, and had a blue marker in one corner of a cube face at the end of one of the "arms" originating from the base square. The task was to decide if the blue marker on the target figure would be positioned above the yellow dot on the reference figure, by rotating and/or folding the target figure to match the reference figure shape. 3 different rotation conditions (0°, 50°, or 150° rotation disparity) and 3 folding conditions (0, 3, or 6 squares to fold) resulted in 9 conditions overall, with the easiest being a visual baseline condition, as it required no spatial transformation. Figure 1 shows an example trial: the reference figure (left) is three-dimensional and partially folded, while the target figure (right) is rotated clockwise by 50 degrees. After rotating and folding the nec-

essary pieces, the blue dot will be positioned above the yellow dot, inducing a "match" response.

Modeling data

Cognitive model A cognitive model was created in ACT-R solving a simulated version of the experiment, consisting of the same time settings, conditions and figure types. It incorporated mental spatial transformation processes suggested by e.g. Just and Carpenter (1976); Yuille and Steiger (1982); Wright et al. (2008). After target onset, the model decides between a direct visual comparison, tries mental rotation, or tries mental folding, mediated by reinforcement learning through rewards for correct answers. Over the course of a trial, it will switch between rotation and folding, depending on the experiment condition. Learning behavior is further simulated by allowing the model to associate shapes to finished transformations and/or trial outcomes, and retrieve these declarative memories at later points in the experiment. The cognitive model made use of a spatial module extension to ACT-R (Preuss et al., 2019; Heimisch et al., 2023), which generalizes mental spatial transformation processes and computes a delay for those processes based on a number of factors. Output from the spatial module is separated into representation and transformation activity, allowing for differentiated analyses of spatial processes. For our example, we focused on the visual, imaginal, retrieval, and spatial modules. The default ACT-R modules were included to serve as an indicator of the sensibility of the resulting matches.

Model fit Once the cognitive model is completed, its behavioral data predictions must be fit as close as possible to the actual experiment results before more complex data comparisons take place. Typically, good model fit is indicated by high correlation and low RMSE after parameter fitting, e.g. by a successive grid search over a sensible range of parameters. The RMSE value can also be used as an indicator of a lag between the model predictions and the observed data, which can be manually corrected (see *Time correction*). The quality of fit can be gauged by performing an ANOVA with data source as a factor – if no significant influence of modeled or human data on RT is found, the goodness of fit should be considered high. Results from our example are shown in Figure 2. In our case, our model was on average slightly slower than human solvers ($M_{exp} = 3.44$, $SD_{exp} = 2.17$; $M_{model} = 4.14$, $SD_{model} = 1.56$), and showed high correlation on group level ($r = .86$, $p < .001$). While human solvers had a success rate of 92.5% ($SD = 26.4\%$), the model reached 100% correctness. An ANOVA as described above showed a significant effect of data source on RT ($F(1, 3171) = 594.6, p < .001$), implying that the model does not fit the data perfectly. However, as average RTs in this experiment and differences between participants were quite high, this discrepancy in fit was considered acceptable.

Data generation After model optimization, activity predictions can be generated. If the order or type of trials presented
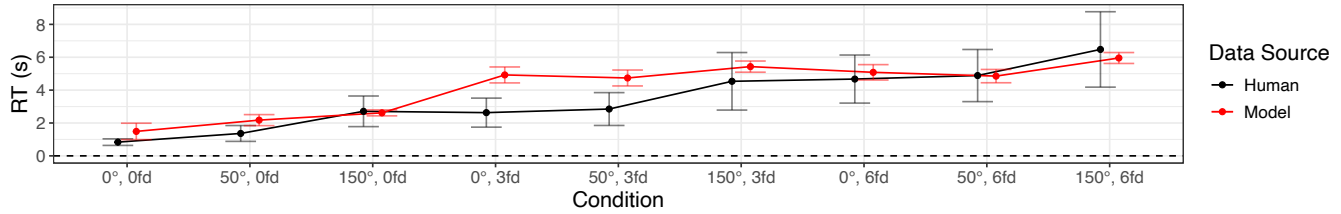
Figure 2: Comparison of RTs from experiment data and model predictions.

to each participant is known, it is possible to increase data comparability further by using individual model instances to simulate specific participants: by following the individual trial order, the simulated activity can be controlled for sequence or learning effects. In addition to modeling the trials of the main experiment, prior training trials, fixation crosses before trials, and breaks between experiment blocks should also be included to control for memory formation and retention effects. The activity of the modules during a cognitive model run should be sampled with the same rate of the final EEG data, which in our case was 250 Hz. Afterwards, the generated data can be aggregated at individual or group level, which for our example was the latter, thereby transforming binary activity data into proportional activity values.

Data smoothing   The model-produced activity could contain a high number of angular features, highly deterministic processes (i.e., always active or non-active) or single-sample spiking, inhibiting its suitability for linear regression. For this reason, a smoothing algorithm can be applied. For our example, we chose kernel smoothing (Wand & Jones, 1994), and picked its parameters to reduce these features while retaining each time series' characteristics as much as possible (first-degree local polynomials, number of grid points equaled sample length, with a bandwidth of 24.). An example of data smoothing on the single conditions of the retrieval module is shown in Figure 3. The end result of the cognitive modeling pipeline will subsequently be referred to as *module activity*.

EEG data

EEG processing   In our experiment, EEG was recorded continuously with 64 active electrodes, including one EOG electrode to identify eye movement artifacts. The electrodes were arranged in the extended 10-20 system and data was recorded with a 500 Hz sample rate, bandpass-filtered from .016 to 250 Hz, and referenced to the FCz electrode. Further pre-processing was applied using the EEGLab toolbox (Delorme & Makeig, 2004), with the BeMoBIL pipeline (Klug et al., 2022). Individual subject data was low-pass filtered at 124 Hz and subsequently downsampled to 250 Hz. ZapLine Plus (Klug & Kloosterman, 2022) was used for frequency noise detection and filtering, resulting in 50 Hz line noise being removed from all datasets. Bad individual channels were interpolated using spherical interpolation and re-referenced to an average reference.
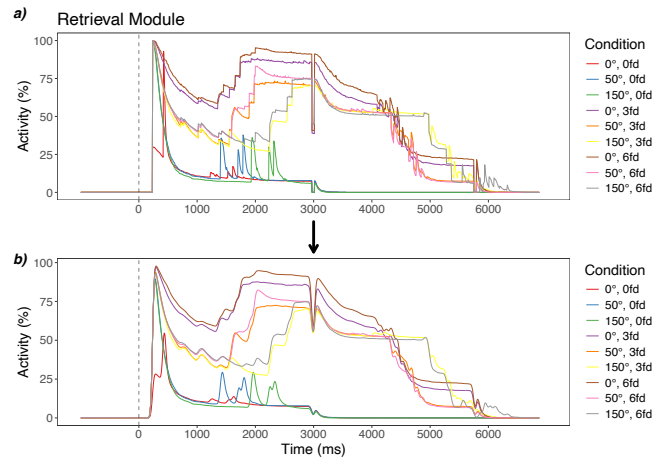

Figure 3: Example of model data smoothing for the retrieval module, separated by condition. Module output (**a**) is smoothed by kernel regression (**b**).

Independent component analysis   Independent component analysis (ICA) is able to separate independent sources that are mixed together across several sensors, including signals originating from the brain and non-brain artifacts. We used an adaptive mixture ICA (AMICA; Palmer et al., 2008) with default time-domain cleaning parameters, and an 1.5 Hz high pass filter, which were both applied only to the AMICA computation and not the final dataset. For each resulting spatial filter, equivalent dipole models were generated (Oostenveld & Oostendorp, 2002). Afterwards, both spatial filters and dipole models were applied to the pre-processed data.

Epoching   A 75 Hz low-pass filter was applied. We selected only components identified by ICLabel (Pion-Tonachini et al., 2019) as brain processes with 70% likelihood. Data streams were then epoched to 200 ms before reference figure onset to 10,000 ms after target onset, and baseline-corrected on the 200 ms pre-reference period. Only "match" trials were retained to filter out implausible reasoning processes. We rejected epochs with RTs beyond 3 median absolute deviations, as well as the 5% noisiest epochs per subject and condition.

IC Clustering (optional)   If analysis on group level is desired, the individual components of each subject that represent activity from a common source region need to be identi-

fied. In our case, we decided on a repeated weighted k-means clustering algorithm. With this approach we ran 10,000 clustering iterations and selected the best solution based on having the highest number of participants with ICs in a specified ROI cluster and a low spread of IC dipole locations within each cluster (see Klug et al. (2022) for a detailed description of the clustering approach and parameters). We chose a target number of 16 clusters, as was used in prior spatial transformation studies (Hilton et al., 2022; Preuss et al., 2023), and specified the central parietal cortex as our ROI (Talairach: x = 0, y = -66.7, z = 41.1), based on an area involved in spatial processing as suggested by a meta-analysis of spatial transformation studies (Zacks, 2008). ICs with dipoles further than 2.5 SD from any cluster centroid were excluded. We then calculated group ERPs for each cluster.

Frequency band filtering (optional) At this point, the cluster ERPs can be further filtered into specific frequency bands. Several studies have focused on specific bandwidths of EEG frequency to focus on specific aspects of cognitive processing, e.g. to analyze spatial processing by alpha-wave activity (e.g. Hilton et al., 2022), or to match cognitive model activity to multiple frequency bands during an attentional blink task (van Vugt, 2012). For this study, no band filtering was applied.

Hilbert transform So far, the cluster ERPs still represent oscillations, i.e. both positive and negative amplitudes from baseline can be considered to be reflective of activity. To achieve a linear signal, the ERPs are Hilbert-transformed. This results in the analytical signal for each ERP, reflecting the non-negative strength of the signal (Le Van Quyen et al., 2001; Kozma et al., 2007). As dissimilar start and end values of the time series can produce artifacts, padding should be added at the beginning and end of the data time series before application of the transform. In our case, we padded each ERP with its reversed first and last 500 ms before and after the signal. After application of the transform, the padding is removed. Figure 4 demonstrates the steps of this process. This final step of the EEG processing pipeline results in what will in the following be referred to as *cluster activity*.

Time correction

Data lag (optional) To get an idea of the specificity of the model fit in time, cross-correlation can be performed. Cross-correlation allows for checking the correlation between two time series in a specific range of sample lag. After individual module output is generated, it can thus be compared to EEG activity to see if lagging the data would result in a better fit on the one hand, and how specific in time a given fit is on the other hand. If need be, it can then be corrected by applying an appropriate lag on module output. Note that as cross-correlation does not account for e.g. autocorrelation or variance, we recommend its use only as a descriptive statistic. In our example, we used cross-correlation to examine possible lag in the ideal correlation of module activity to cluster
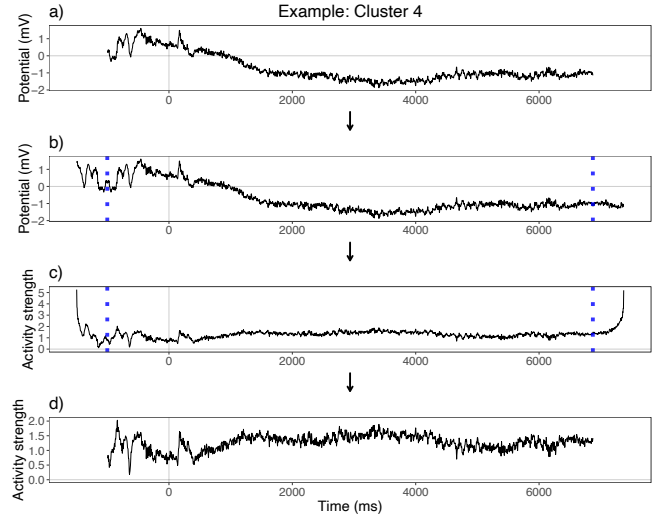

Figure 4: Example of Hilbert transforming a cluster ERP. The original ERP **(a)** is padded by adding its reversed first and final values **(b)**. After applying Hilbert transform to receive the absolute analytical signal **(c)**, the padding is removed **(d)**.

activity, especially for the spatial module, which is shown in Figure 5. We explored a range of ± 500 ms. The representational output shows a good, determinate fit with multiple clusters, including the ROI cluster. At the same time, a drift for parietal clusters 5 and 16 by 40 (160 ms) respectively 60 samples (240 ms) is apparent, implying representation output to be too slow to fit well to these clusters. Transformational output shows a good but indeterminate fit to the ROI cluster, as well as to most other clusters. Based on these results, we chose not to apply lag correction.

Time warping (optional) If the fit in response times between the experiment data and model-generated output during individual trials is not close enough, some authors suggest time warping of model activity to match human RTs (Anderson et al., 2008; Borst et al., 2011). While this increases comparability of overall model activity, it might falsify the individual predictions of the involved processes. If no time warping is applied, the shorter dataset should instead be zero-filled until the maximum trial length to facilitate averaging of trials. We chose not to perform time warping on our generated data, as we were especially interested in specific intra-trial processes and their individual predictive quality, instead of overall activity during trials.

GLS regression

Ideally, both module and cluster activity match in their post-processing sample rate. If not, one of the datasets needs to be adjusted by down-sampling, in which case we recommend resampling module activity as it has a lower information rate per sample. For our datasets, both sample rates matched at 250 Hz. Furthermore, module and cluster activity should be limited to the respective time range of activity of the module
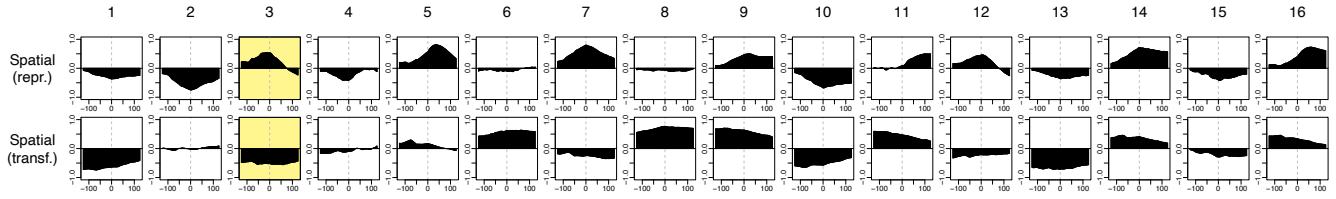
Figure 5: Cross-correlation for the representational and the transformational spatial module output. ROI cluster highlighted.

(i.e., its first and last sampled activity) to avoid zero inflation in comparing the two datasets.

Finally, group-level GLS linear regression for each module is performed with all available clusters as predictors, and added structures to control for autoregression of predictors and variance differences in conditions. It is possible to use GLS regression over all data, or group it into conditions. In our case, we used first-order autocorrelation (Box et al., 2015) and scaled identity variance structures (Carroll & Ruppert, 2017), grouped by the 9 difficulty conditions. Any cluster that forms a significant coefficient is considered a match for that specific module, with the significance threshold set at $p = .001$. Goodness of fit of each GLS linear model is then determined by comparing its likelihood ratio with that of a null model, and testing the significance of the improvement.

Results

For each module, matches between module and cluster activity were identified. The clusters found to be significant now form a *cluster structure*, which are associated with the function of that module and should form a cohesive brain area. The clusters with the best fit resulting from our example analysis, judged by the strength of their coefficients in the GLS regressions, are shown in Figure 6, while Figure 7 displays the resulting cluster structures for each selected module. Note that while the imaginal and retrieval module share a right parietal cluster as the best match for activity, both differ in their final cluster structures. The visual structure consists of a single, superior parietal cluster. For imaginal activity, a structure of parietal, right parietal, and right frontal clusters was identified. Retrieval activity is matched by right, central, and left parietal, as well as left and right frontal clusters. Activity reflecting spatial representation formation is reflected by central, left, and right parietal clusters. Finally, spatial transformative activity matches a central and left parietal cluster structure.

Discussion

By applying the analysis outlined in this paper on example data from a cognitive model and EEG recordings from a combined mental rotation and folding task, we were able to deduce cohesive, sensible structures from clustered EEG activity that were matched to simulated activity produced by distinct cognitive processing units during trials. In our example, our goal was to identify neural substrates for spatial activity, for which we were able to uncover parietal areas matching

our model predictions, as well as results from several prior studies (e.g., Harris et al., 2000; Zacks, 2008). Moreover, the resulting structures show differentiation between representative and transformative activity, affirming previous studies that have found spatial storage and manipulation activity to differ in lateralization (Milivojevic et al., 2003; Gardony et al., 2017; Hilton et al., 2022; Preuss et al., 2023). Our additional analyses of non-spatial modules could be considered a sanity check (anchoring an exploratory search for appropriate brain areas on known module-brain associations), and yielded similar results to Borst and Anderson (2013), albeit seemingly biased towards spatial processing (e.g., identifying no occipital clusters for the visual module). With regard to these results, we believe our approach to be ideally suited for finding neural correlates of cognitive activity without *a priori* established locations.

The application of this approach opens up additional possibilities. While our example used data from an ACT-R model and EEG recordings, the steps outlined herein should largely hold true for data from other cognitive architectures and/or neuroimaging techniques. A basic requirement for this are event-matched time series from both simulated and recorded experiment output, sampled with a high enough frequency to detect features of interest. Furthermore, instead of averaging over all conditions, an intriguing use case lies in the application of this method separately for contrasting conditions, e.g. baseline vs. target, and comparing the respective structures with each other. Another possibility to further refine the outcome manifests in comparing module and cluster activity in the time-frequency domain: in the case of certain cognitive functions or brain areas, different frequency bands are associated with distinct functions. Exploration of data in this manner could help, and may indeed be necessary, to reach more fine-grained conclusions. Finally, the resulting cluster structures lend themselves to additional analysis. For instance, the functional influence of the structures on one another could be investigated by connectivity analysis. Used as an exploratory tool, brain areas suggested by this method could be considered ROIs in further studies, giving initial insight into hitherto unexplored, specific cognitive functions.

Of course, the methodology comes with several caveats. Linear regression with GLS, including correlation and variance structures, is computationally heavy: for our example data, fitting a single GLS model took 5 orders of magnitude longer than a regular linear model on mid-range consumer hardware. High CPU and RAM resources are recommended.
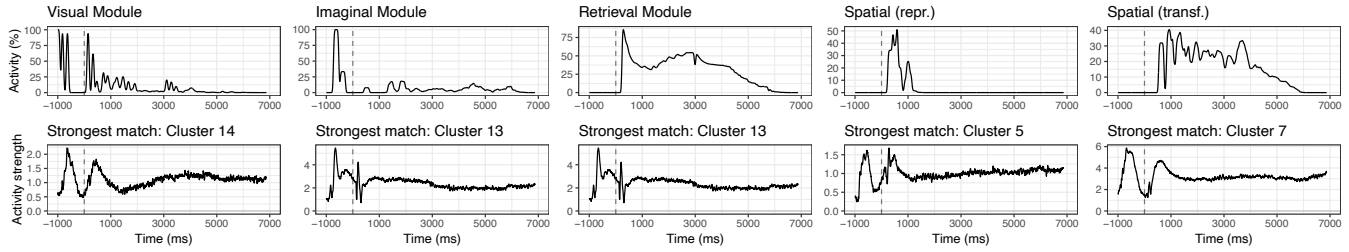
Figure 6: Best fitting clusters for five selected modules. First row: module activity, averaged over all conditions. Second row: cluster activity of the strongest predictive quality for each module, averaged over all conditions.
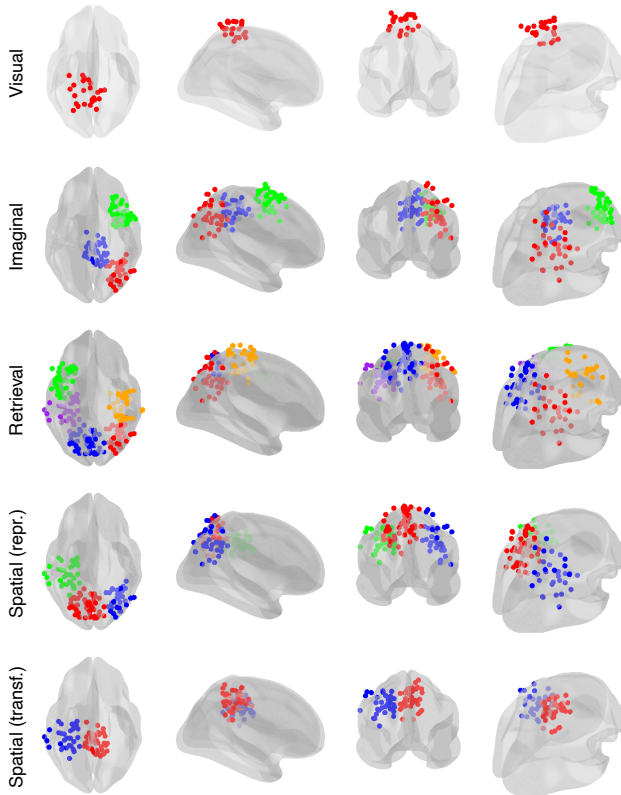

Figure 7: Brain renderings showing cluster structures comprised of every matching cluster per module.

While the cluster structures uncovered in our example are largely reasonable, the method essentially identifies matching patterns between the simulated and recorded datasets, and does not explicitly signify functional matches. In addition, individual clusters with a significant predictivity of module activity are not necessarily exclusive to a single module, but can be shared among multiple structures. In our case, a right parietal cluster reached significance for three different module GLS regressions. Depending on the usage scenario, this could add complexity to the understanding of the results, or further analyses thereof. Sensible interpretation of the produced structures is therefore especially advisable.

At the time of writing, multiple upcoming studies make use of the methodology presented in this paper, including results

from separate mental rotation and mental folding experiments (Preuss et al., 2023). Further work on the cMRF dataset will extend the approach with a subsequent connectivity analysis, focusing on the influence of each cluster structure on each other cluster structure.

To summarize, we have presented a method that facilitates finding neural substrates of simulated activity produced by cognitive models, and which is especially suited for exploratory work on specific cognitive functionality. Applied on data from a mental spatial transformation study, we have found spatial activity in several parietal areas that could be differentiated in representational and transformational cluster structures. On a final note, we do consider the development of this method an ongoing process, and would welcome feedback and improvements on it.

Acknowledgments

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036.

Anderson, J. R., Carter, C. S., Fincham, J. M., Qin, Y., Ravizza, S. M., & Rosenberg-Lee, M. (2008). Using fMRI to test models of complex cognition. *Cognitive Science*, *32*(8), 1323–1348.

Borst, J. P., & Anderson, J. R. (2013). Using model-based functional mri to locate working memory updates and declarative memory retrievals in the fronto-parietal network. *Proceedings of the National Academy of Sciences*, *110*(5), 1628–1633.

Borst, J. P., & Anderson, J. R. (2017). A step-by-step tutorial on using the cognitive architecture ACT-R in combination with fMRI data. *Journal of Mathematical Psychology*, *76*, 94–103.

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2011). Using a symbolic process model as input for model-based fMRI analysis: Locating the neural correlates of problem state replacements. *NeuroImage*, *58*(1), 137–147.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Carroll, R. J., & Ruppert, D. (2017). *Transformation and weighting in regression*. Chapman and Hall/CRC.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9–21.

Gardony, A. L., Eddy, M. D., Brunyé, T. T., & Taylor, H. A. (2017). Cognitive strategies in the mental rotation task revealed by EEG spectral power. *Brain and Cognition*, *118*, 1–18.

Harris, I. M., Egan, G. F., Sonkkila, C., Tochon-Danguy, H. J., Paxinos, G., & Watson, J. D. (2000). Selective right parietal lobe activation during mental rotation: a parametric PET study. *Brain*, *123*(1), 65–73.

Heimisch, L., Preuss, K., & Russwinkel, N. (2023). Cognitive processing stages in mental rotation – how can cognitive modelling inform HsMM-EEG models? *Neuropsychologia*, *188*, 108615. doi: https://doi.org/10.1016/j.neuropsychologia.2023.108615

Hilton, C., Raddatz, L., & Gramann, K. (2022). A general spatial transformation process? Assessing the neurophysiological evidence on the similarity of mental rotation and folding. *Neuroimage: Reports*, *2*(2), 100092.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480.

Katanoda, K., Matsuda, Y., & Sugishita, M. (2002). A spatio-temporal regression model for the analysis of functional MRI data. *NeuroImage*, *17*(3), 1415–1428.

Klug, M., Jeung, S., Wunderlich, A., Gehrke, L., Protzak, J., Djebbara, Z., ... Gramann, K. (2022). The BeMo-BIL pipeline for automated analyses of multimodal mobile brain and body imaging data. *bioRxiv*, *09*.

Klug, M., & Kloosterman, N. A. (2022). Zapline-plus: A Zapline extension for automatic and adaptive removal of frequency-specific noise artifacts in M/EEG. *Human Brain Mapping*, *43*(9), 2743–2758.

Kozma, R., Aghazarian, H., Huntsberger, T., Tunstel, E., & Freeman, W. J. (2007). Computational aspects of cognition and consciousness in intelligent devices. *IEEE Computational Intelligence Magazine*, *2*(3), 53–64.

Lawson, C. L., & Hanson, R. J. (1995). *Solving least squares problems*. SIAM.

Le Van Quyen, M., Foucher, J., Lachaux, J.-P., Rodriguez, E., Lutz, A., Martinerie, J., & Varela, F. J. (2001). Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony. *Journal of neuroscience methods*, *111*(2), 83–98.

Milivojevic, B., Johnson, B., Hamm, J. P., & Corballis, M. C. (2003). Non-identical neural mechanisms for two types of mental transformation: event-related potentials during mental rotation and mental paper folding. *Neuropsychologia*, *41*(10), 1345-1356. doi:

https://doi.org/10.1016/S0028-3932(03)00060-5

Oostenveld, R., & Oostendorp, T. F. (2002). Validating the boundary element method for forward and inverse EEG computations in the presence of a hole in the skull. *Human brain mapping*, *17*(3), 179–192.

Palmer, J. A., Makeig, S., Kreutz-Delgado, K., & Rao, B. D. (2008). Newton method for the ICA mixture model. In *2008 IEEE international conference on acoustics, speech and signal processing* (pp. 1805–1808).

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181-197. doi: https://doi.org/10.1016/j.neuroimage.2019.05.026

Preuss, K., Hilton, C., Gramann, K., & Russwinkel, N. (2023). *Finding common ground for mental spatial transformation - cognitive models identify neural substrates of mental rotation and folding*. Berlin, Germany. (Manuscript submitted for publication)

Preuss, K., Raddatz, L., & Russwinkel, N. (2019). An implementation of universal spatial transformative cognition in ACT-R. In T. D. Stewart (Ed.), *Proceedings of the 17th international conference on cognitive modelling* (pp. 144–150). Waterloo, Canada: University of Waterloo.

Prezenski, S., & Russwinkel, N. (2016). A proposed method of matching ACT-R and EEG-data. In D. Reitter & F. Ritter (Eds.), *Proceedings of the 14th International Conference on Cognitive Modeling* (pp. 249–251). PA: Penn State: University Park.

Sato, J. R., Junior, E. A., Takahashi, D. Y., de Maria Felix, M., Brammer, M. J., & Morettin, P. A. (2006). A method to produce evolving functional connectivity maps during the course of an fMRI experiment using wavelet-based time-varying Granger causality. *Neuroimage*, *31*(1), 187–196.

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology*, *3*(2), 228–243.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

van Vugt, M. K. (2012). Relating ACT-R buffer activation to EEG activity during an attentional blink task. In N. Russwinkel, U. Drewitz, & H. van Rijn (Eds.), *Proceedings of the 11th international conference on cognitive modelling* (p. 218).

Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. Chapman & Hall.

Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin & Review*, *15*(4), 763–771.

Yuille, J. C., & Steiger, J. H. (1982). Nonholistic processing in mental rotation: Some suggestive evidence. *Perception & Psychophysics*, *31*(3), 201–209.

Zacks, J. M. (2008). Neuroimaging studies of mental rotation: a meta-analysis and review. *Journal of cognitive Neuroscience*, *20*(1), 1–19.

Understanding Human Behavior and a Cognitive Model in an Image Labeling Task

Jongchan Pyeon
(jbp5681@psu.edu)
Department of Industrial and
Manufacturing Engineering,
The Pennsylvania State University
University Park, PA 16802, USA

Amir Bagherzadeh
(amir.bagherzadeh@psu.edu)
Department of Industrial and
Manufacturing Engineering,
The Pennsylvania State University
University Park, PA 16802, USA

Roya Koshani
(rvk5607@psu.edu)
Department of Chemical
Engineering,
The Pennsylvania State University
University Park, PA 16802, USA

Amir Sheikhi (sheikhi@psu.edu)
Departments of Chemical Engineering, Biomedical
Engineering, and Neurosurgery
The Pennsylvania State University
University Park, PA 16802, USA

Farnaz Tehranchi (farnaz.tehranchi@psu.edu)
School of Engineering Design and Innovation,
The Pennsylvania State University
University Park, PA 16802, USA

Abstract

The field of Artificial Intelligence (AI), particularly in the area of computer vision, has experienced significant advancements since the emergence of deep learning models trained on extensively large labeled datasets. However, reliance on human labelers raises concerns regarding bias, inconsistency, and ethical issues. This study aimed to replace human labelers with an interactive cognitive model that could address these concerns. We investigated human behavior in a two-phase image labeling task and developed a model using the VisiTor (Vision + Motor) framework within the ACT-R cognitive architecture. This study was designed based on a real labeling task of identifying different crystals in optical microscopy images after various treatments for inhibiting the formation of the crystals. The outcomes from the image labeling experiment, which included both learning and testing phases, revealed meaningful observations. The observed decrease in task completion times for all participants during the learning phase suggests an increased familiarity with the image features, facilitated by the reference images presented in all four consecutive example tasks. It was also discovered that the subtle distinctions between classes led to confusion in making decisions about labels. The developed interactive cognitive model was able to simulate human behavior in the same labeling task environment, while the model achieved high accuracy, it still relies on pre-defined features therefore limited its application to seen data only. Our findings suggest that interactive cognitive modeling offers a promising avenue for replacing human labelers with robust, consistent, and unbiased labeled datasets.

Keywords: Object labeling; Cognitive modeling; ACT-R; Learning

Introduction

Artificial Intelligence (AI) advances have been significantly boosted since the publication of 'ImageNet Classification with Deep Convolutional Neural Networks' (Krizhevsky et al., 2012). In the fields of computer vision and deep learning, such as with Convolutional Neural Networks (CNNs) and Vision Transformer (ViT), utilizing datasets consisting of a vast number of labeled images, such as ImageNet (Deng et al., 2009), is essential for building models for image classification, image segmentation, or object detection. Therefore, human labelers are tasked with reviewing and annotating data, such as images, text, or videos, by assigning labels and tags that will be used for training these models. In general, crowd-sourcing services, such as Amazon's Mechanical Turk (MTurk), have been widely used for image annotation, thanks to their relatively low cost and quick turnaround (Rashtchian et al., 2010).

Concerns have arisen regarding human labelers in annotation tasks despite mentioned progress. The presence of labeler bias in datasets can lead to inaccurate or unfair decisions in various fields, including law enforcement, healthcare, and education, due to the labelers who hold stereotypes based on different ethnicities and genders (Haliburton et al., 2023). Moreover, labeling inconsistencies resulting from labelers' inherent judgments and other factors can lead to significant errors in decision-making, particularly in clinical settings (Sylolypavan et al., 2023). Furthermore, ethical concerns have been raised regarding the use of humans as labelers. Several social issues related to manual image labeling work performed by humans have been divulged as a result of its low pay, poor working conditions, and psychological repercussions (Hagendorff, 2020).

Therefore, there is a need to replace human labelers with models or at least part of the labeling process. The best candidate for this role would be a cognitive model, as cognitive models can serve as surrogate labelers. However, building a robust cognitive model capable of taking over the role of humans as labelers presents a significant challenge. It is imperative to understand human behavior and predict human performance in order to address this challenge, which has remained a long-standing issue for researchers. Nevertheless, this endeavor will help us gain insights into why people perform better in certain situations and why they may not perform as well in others.

The cutting-edge method of simulating behavior involves the use of cognitive architectures (Kotseruba et al., 2016). These architectures are capable of replicating the cognitive processes that the human mind undergoes to complete tasks. By accurately modeling the human cognitive process using cognitive architectures, researchers can predict completion times for actions such as eye movements and motor actions (i.e., physical actions). Additionally, cognitive modeling of human behavior enhances our understanding of cognitive

processes across various domains. Unlike data-analytical models, these cognitive models provide a deeper theoretical insight into cognitive functions and how the mind works (Pinker, 2003).

Newell (1990) introduced the concept of unified theories of cognition, which involves acquiring knowledge, problem-solving, and perception. Cognitive architectures, such as Adaptive Control of Thoughts-Rational (ACT-R) (Anderson et al., 1998; Anderson & Lebiere, 2014; Ritter et al., 2018) and SOAR (Laird, 2019), have been developed to model human cognition in various cognitive tasks. ACT-R can interact with the environment through modules like perception (vision) and action (motor) modules. This mechanism enables the integration of human cognition to simulate human cognition and behavior (Tehranchi & Ritter, 2017, 2018).

ACT-R is limited interaction ability with dynamic environments because it interacts with environments either constructed within ACT-R's simulated Lisp environment and facilitated through its device interface (Byrne, 2001). Hence, we use the VisiTor framework to make ACT-R interact with dynamic task environments, such as the experiment environment. VisiTor (Vision + Motor) facilitates ACT-R's interaction with diverse environments through two main modules: Vision and Motor, so that participants' visual attention and mouse control can be simulated (Bagherzadeh & Tehranchi, 2022; Bagherzadehkhorasani & Tehranchi, 2023).

In this study, we investigate the behavior of labelers and the capabilities of a developed interactive cognitive model for simulating and understanding human behavior in image labeling tasks. Our aim is to understand how labelers behave in the image labeling task. To achieve this, we designed the task with two phases: learning and testing. Throughout the study, we collected participants' interaction data via a research-grade eye-tracking device to examine their visual attention on the presented objects. Understanding human behavior in this task is crucial for building a cognitive model capable of simulating human cognition. Furthermore, we explore the capability of VisiTor to simulate human behavior in the same image labeling task. We use only the testing phase data for developing a cognitive model using the VisiTor framework.

Methodology

This study aims to understand human behavior in an image labeling task. Also, the task is designed to observe participants' visual attention using an eye-tracker while they navigate the environment. The task is designed in a Google Forms. Participants can interact with the form and select items. We used VisiTor to simulate participants' behavior in the same task environment that participants used (i.e., Google Forms) that is a dynamic task environment. This approach allows us to develop an interactive cognitive model using the ACT-R. This study was approved by the Institutional Review Board of The Pennsylvania State University (IRB approval number: STUDY00024434).

Task Environment Design

The form for the image labeling study consists of two phases: the learning phase and the testing phase.

In the learning phase, we plan to examine what features of the objects in each class are visually explored by the participants. Therefore, this phase is composed of two stages. The first stage of the learning provides reference images for each of the four classes. Each reference image consists of four objects, cropped from the optical microscopy image, with each object corresponding to a similar shape within its respective class, as shown in Figure 1 (a) and (b). We do not provide any descriptions of the shape features of each class, except for class 4, because it is an object belonging to neither Class 1, 2, nor 3. The second stage of the learning is composed of four example tasks, one for each class. The answers with the reference images for each class are given in all the examples to help participants understand what the main tasks look like and how they can answer the questions in the testing phase.
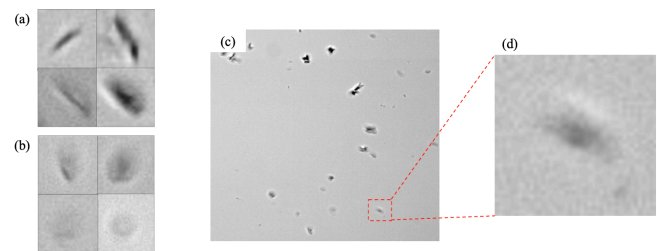
Figure 1: (a) The reference images for Class 1. (b) The reference images for Class 4. (c) A portion of the original microscopy image showing the target object of Task 14. (d) A magnified view of the target object.

In the testing phase, participants will encounter a total of 20 tasks, distributed evenly across four classes, resulting in 5 tasks per class. Each task will feature an image containing multiple objects, with one object of interest highlighted by a red bounding box. To aid in the examination of this object, a magnified view of the area within the bounding box will also be provided, ensuring participants can closely analyze the target object, as shown in Figure 2. Unlike the examples provided in the learning phase, the reference images are not provided in the testing phase. Therefore, participants will have to rely on their memory.

Image Dataset The optical microscopy image was captured during an experiment investigating the effect of various treatments on inhibiting the formation of calcium oxalate crystals. Therefore, the objects appearing in the image are crystals, which may be various forms of calcium oxalate crystals or other undefined crystals. Each class—1, 2, and 3—corresponds to COM-I (individual calcium oxalate monohydrate), COM-aggregates (aggregated calcium oxalate monohydrate), and COD (calcium oxalate dihydrate), respectively. These crystals were labeled by an expert in the field. Each object used for the reference images, examples,

and main tasks, was cropped from the original microscopy image.

Figure 2: The image labeling task Google Forms. Left image is an image with multiple objects including a target in the red bounding box. Right image is a magnified view of the area withing the bounding box.

Eye-tracking Data Collection

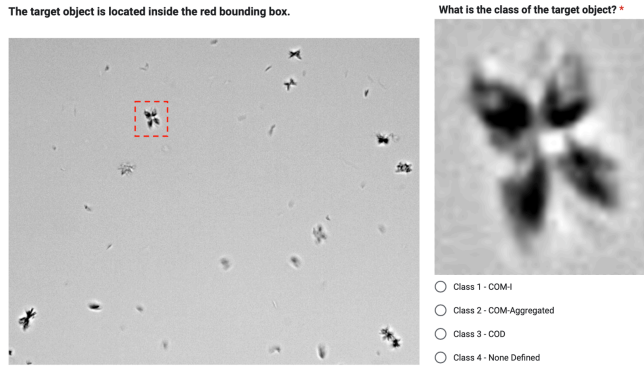The participants' eye movements were collected by the *GazePoint GP3 HD* eye tracker, as shown in Figure 3(a). The device was positioned directly below the computer monitor at approximately 65 cm from the participants' eyes, to ensure accurate tracking of eye movements during the study.

The visualization of these movements was achieved through a *Fixation Map* in *GazePoint Analysis* 6.11.0, as illustrated in Figure 3(b). Two computer monitors were utilized: one monitor recorded participants' eye movements using the *GazePoint Analysis* software provided by the manufacturer, while the other was designated for participants to perform the image labeling task, as illustrated in Figure 4. This setup allowed for the tracking and collection of visual attention data based on the participants' eye movements.
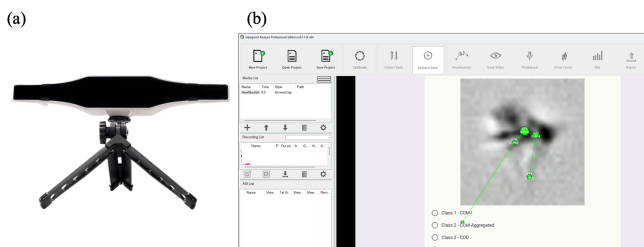

Figure 3: (a) *GazePoint GP3 HD* eye-tracker device. (b) The eye movement recording process.

Description of the eye-tracker The *GazePoint GP3 HD* eye tracker is a research-grade eye tracker. It has a sampling rate of 60 to 150 Hz, meaning that eye movements are recorded as frequently as every 6.7 ms at 150 Hz, and a visual angle accuracy of 0.5° to 1°.

User Study Procedure

Participants were welcomed to sit in a chair in front of the computer screen, adjusting their position to comfortably reach the mouse and ensure their eyes were within the eye-

tracking device's detection range. They were then briefed on the study's purpose and procedure and asked to read the instructions that appeared throughout the study carefully. Before the learning phase began, the eye tracking calibration process was implemented for each participant to optimize eye movement tracking. After the calibration process was completed, and the recording of eye movements began, participants transitioned into the learning phase. During this phase, they were given ample time to familiarize themselves with the features of each class through the two stages: the reference images and the four example tasks. Having completed these preparatory stages, participants then moved on to the testing phase, where they worked on the image labeling tasks based on what they had observed in the learning phase. After they completed all the 20 tasks, they were asked to submit their answers, after the 20th question the recording was stopped.
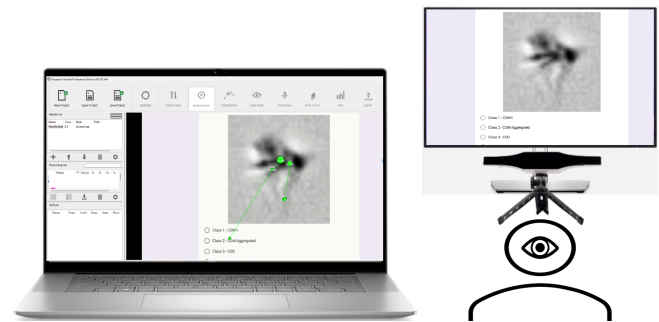
Figure 4: User study setup: The screen on the right with the eye-tracker is where a participant performs tasks, and the participant's eye movements are recorded on the left screen.

Demographics The participants (*n* = 5, Female: 2) are graduate students and faculty members from the Pennsylvania State University.

Interactive Cognitive Modeling Approach

An interactive cognitive model for simulating human behavior in the image labeling task was developed using VisiTor. VisiTor consists of two main modules: Vision and Motor. These modules work alongside (or enable) cognitive architectures in dynamic environments to perform tasks. The *Vision* module identifies a matched object in a given environment through a template matching pipeline using the OpenCV library, while the *Motor* module executes desired manual actions (e.g., moving a cursor, clicking) through using PyAutoGUI library in python (Bagherzadeh & Tehranchi, 2022). To build an interactive cognitive model in ACT-R, it is necessary to define both declarative knowledge (chunks) and procedural knowledge (production rules). The model was specifically designed to perform tasks exclusively during the testing phase. This emphasis on testing phase performance highlights the experimental nature of the model, particularly in its application to the image labeling task with the VisiTor framework. Also, we used the default learning

parameters within the ACT-R that are recommended by the ACT-R user manual (Bothell, 2017).

We began with declarative knowledge, wherein attributes and their corresponding values were defined for each visual object such as target object images. From the original optical microscopy image, which contained various objects, a total of 20 target object images were cropped, resulting in 5 target object images for each class. The semantic meanings defined for each class in declarative knowledge, rather than the specific chunks for each object image, were used to determine the class of each target object image. Additionally, images from the image labeling task (i.e., the form) that contained 'Class' options and a 'Next' button were cropped, and chunks were defined for these User Interface (UI) elements. Following this, production rules were established to execute the *Vision* module in ACT-R for shifting visual attention and the *Motor* module in ACT-R for actions including clicking, pressing a key, and moving the cursor.

The flowchart detailing the defined procedural knowledge for the task is illustrated in Figure 5. The task initiation is represented by the examination of the first cropped target object image. The *Vision* module then shifts attention based on the cropped image's defined patterns to locate the target object within the larger, original microscopy image that contains various objects. Subsequently, the *Motor* module is engaged to press the 'End' key, scrolling the form to the section where a multiple-choice option is located. Utilizing the pattern matching algorithm, the model shifts its attention to the correct choice among four options. At this stage, the *Motor* module moves the mouse cursor to the identified option and selects it. The process continues with the model finding the 'Next' button through pattern matching, shifting attention to this button, moving the cursor to it, and executing a click. These steps are repeated for each test until the task is completed. A recursive loop, indicating this repetition until the task's conclusion, is represented by a blue arrow in Figure 5.
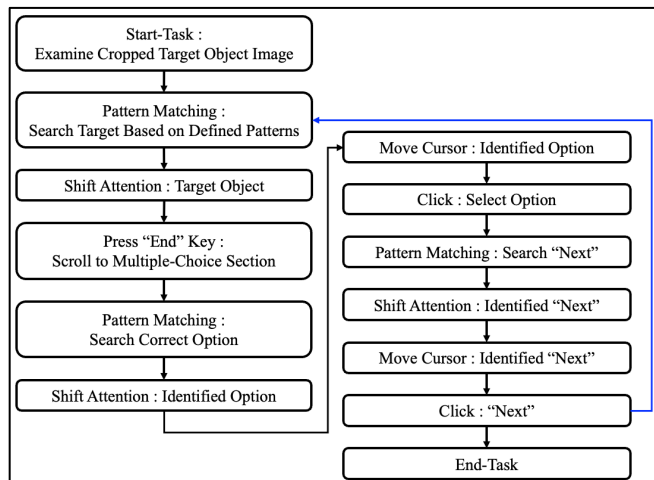

Figure 5: Flowchart representing the implementation of production rules in the interactive cognitive model developed with the VisiTor framework within the ACT-R.

Results

Task completion time was measured for each participant (n = 5) during both the learning and testing phases. This measurement was used instead of labeling accuracy to analyze the participants' behavior in each phase because all participants provided the same incorrect answer. In addition, an interactive cognitive model, constructed within the ACT-R using the default learning parameters, was developed to simulate human behavior and predict completion time in this image labeling task using the VisiTor framework (Bagherzadeh & Tehranchi, 2022).

Learning Phase

We measured how much time the participants spent on reference images and each example. All the measured times are provided in Table 1. The completion time for the reference images was measured by counting the time frames after a participant moved to the Reference section page. Since participants scrolled up and down during the Reference section, we decided not to calculate the time spent on each class. Instead, we measured the total time spent in the entire Reference section. Similarly, the completion time for the Example section was measured by counting the time frames each participant spent on each example. However, in calculating the total time spent in the learning phase, we excluded the time participants spent reading the instructions that appeared before the beginning of the Example section to navigate participants to the new section, as well as the time taken for Google Forms to reload its content. We observed that the participants spent more than 30 percent and up to 55 percent of their total time in the learning phase. Additionally, a trend was discovered where the time spent by participants on each example decreases as they progress through the examples.

Table 1: Participants' task completion time during the learning phase for each example.

Participant	Reference images	Example 1 (Class 1)	Example 2 (Class 2)	Example 3 (Class 3)	Example 4 (Class 4)
P1	36s	12s	9s	10s	6s
P2	20s	11s	7s	10s	5s
P3	18s	15s	10s	10s	8s
P4	42s	21s	15s	5s	3s
P5	30s	11s	6s	4s	4s

Testing Phase

Task completion time for each participant was measured by calculating the time frames from the moment the task contents were fully reloaded to when the 'Next' button was clicked. On average, each participant experienced a 1-second delay in content reloading after clicking 'Next' button. Table 2 displays the average completion time for each class, calculated by dividing the total completion time of tasks for each class by five, to account for the five tasks per class. Additionally, the 'Total' column in Table 2 presents the calculated average task completion time for each task during the testing phase for each participant.

Table 2: Each 'Class' column represents the average completion time per class and 'Total' column represents the total average completion time during the testing phase.

Participant	Class 1	Class 2	Class 3	Class 4	Total
P1	4.2s	3.8s	6.8s	2.4s	4.30s
P2	4.0s	5.0s	4.8s	4.0s	4.45s
P3	6.0s	7.0s	7.8s	5.4s	6.55s
P4	3.2s	5.0s	5.4s	2.8s	4.10s
P5	2.8s	3.4s	6.6s	2.8s	3.90s

In most tasks, all participants completed tasks around their average task completion time per class. However, in Task 5 (for Class 3), for both participants 1 and 5, outliers with significantly longer task completion times—17 and 19 seconds, respectively—were observed compared to their average task completion times of 4.3 and 3.9 seconds, respectively. This discrepancy in completion times, though notable, did not affect the accuracy of their responses, as both participants 1 and 5 selected the correct answer for Task 5. In the case of Participant 5, after initially selecting 'Class 2' as the answer for Task 5, he moved on to Task 6. However, after briefly working on Task 6, he returned to Task 5 and changed his answer to 'Class 3'. The returning time was also included in the Task 5 completion time.
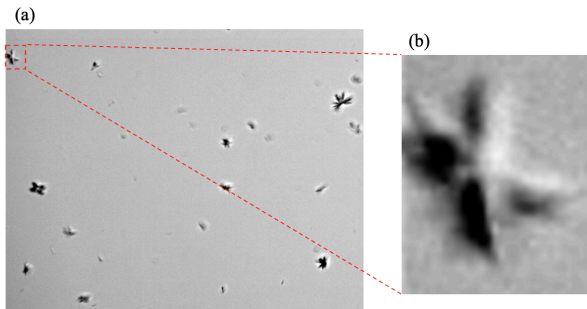

Figure 6: (a) A portion of the original microscopy image showing the target object of Task 5. (b) A magnified view of the target object.

Moreover, for Task 14, all five participants selected the same incorrect answer. Figure 1 (c) and (d) display the images presented in Task 14. The correct answer, 'Class 1', is illustrated in Figure 1 (a). However, they submitted 'Class 4', as shown in Figure 1 (b), as their answer. Hence, they all attained the same score, 19 out of 20, in the image labeling task.

Interactive Cognitive Model Implementation

The human cognition process in a dynamic environment, specifically the image labeling task, was simulated by the ACT-R model. The *Vision* and *Motor* modules introduced by VisiTor enabled the development of an interactive cognitive model tailored to this task. Based on the designed flowchart for the image labeling task, as depicted in Figure 5, we successfully crafted and executed the interactive cognitive model to simulate human behavior in the task.

Figure 7 illustrates the model's functionality during the testing phase of the image labeling task. Figure 7 (a)

showcases the 'Shifting attention' capability of the *Vision* module, showing its process for identifying a matching object within the image through a template. This action is represented by positioning the eye icon near the red bounding box. Subsequently, Figure 7 (b) reveals the outcome of triggering the 'Press "End" key' operation, succeeded by 'Shifting attention' again to pinpoint the correct choice using the model's pre-set memory chunks. Figures 7 (c) and (d) demonstrate the successful sequential execution of production rules: 'Moving the cursor', 'Clicking "Class 2"', 'Shifting attention', 'Moving the cursor', and 'Clicking "Next"', in that order. The model's implementation in the actual image labeling task, as depicted in Figure 7, indicates that the developed model is capable of simulating human behavior within the task in a dynamic environment. This environment reflects the actual actions of the participants during the task, including movements of the mouse cursor, page scrolling, and updates to content on the page.
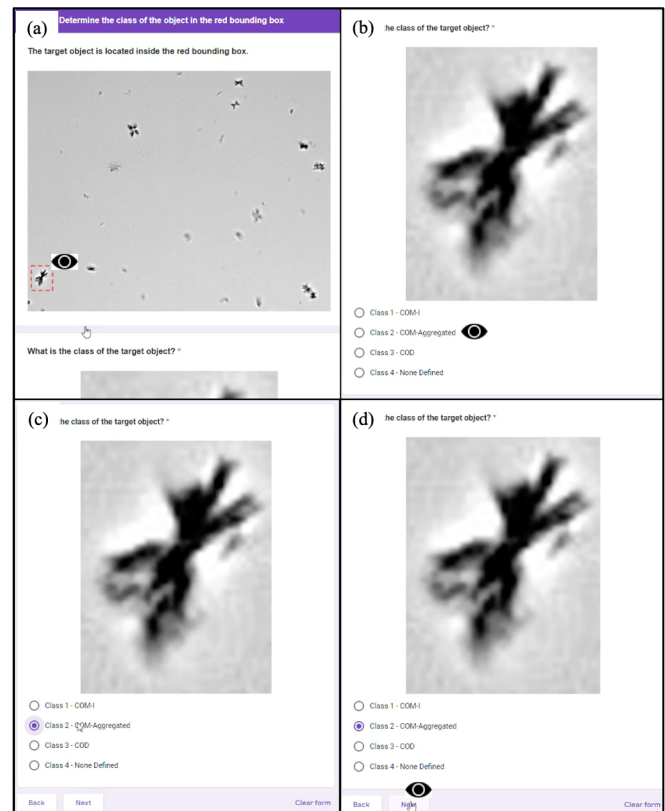
Figure 7: Sequential demonstration of the developed interactive model's cognitive processes during the testing phase (The eye icon represents where the model looks).

However, as mentioned in the Methodology section, the learning parameters within the ACT-R model were not modified to optimize the model's memory retrieval performance. Consequently, the completion time for nearly all tasks and the average completion time across classes were consistently 0.805 seconds, with the exceptions of Task 20 and Class 1, as illustrated in Figure 8. A completion time of 0.755 is because Task 20 is the last task so there is no need

for firing procedural rules for preparing the next batch for starting the new task.
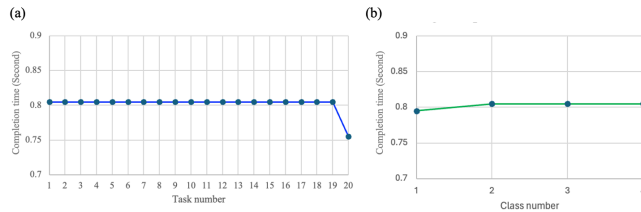
Figure 8: (a) The model completion time plotted for each task. (b) Averaged model's completion time plotted for each class.

Discussion

The observed outcomes from the behavioral study of humans in the image labeling task, along with the development of an interactive cognitive model for simulating human behavior in the same task, present many novel findings. This is despite the study being an initial labeling study with simple but real labeling tasks.

The findings from the image labeling experiment, which consisted of two phases, uncover human behavior and suggest limitations. In the learning phase, the observed trend of decreasing completion times as the participants progressed through the example tasks implies that the participants might study the characteristic patterns of each class not only from the Reference section but also from the Example section, given that all the example tasks include the reference images for each class. Consequently, the participants spent the least time on Example 4, having learned the features of each class as they completed the preceding tasks. During the testing phase, several interesting findings emerged. Notably, participants 1 and 5 exhibited significantly longer completion times for Task 5, likely due to shaded zones around the target object. Figure 6 (b) reveals that these shades, caused by the protruding object and optical microscopy lighting, initially confused the participants. However, once they recognized that the black areas were merely shadows, both participants successfully identified the target as belonging to Class 3. This insight explains the additional time they needed before reaching the correct conclusion. Additionally, all participants selected the same incorrect answer for Task 14. As illustrated in Figures 1 (a), (b), and (c), the distinction between Class 1 and Class 4 is subtle in this case. This uniform error may be attributed to the insufficient provision of reference images, which restricted the participants' capacity to adequately learn and differentiate the characteristic patterns of each class. Moreover, the study's depth into human behavior was limited by the small number of participants and tasks, preventing the derivation of substantial insights. The simplicity of object shapes and the scant variety of classes further narrowed the scope of our observations. Additionally, the eye-tracking device's accuracy fell short in tracking eye movements across edges or lines crucial for label determination. Therefore, we utilized the eye-tracking data to observe where the participants looked and for how long (i.e., shifting attention).

We developed an interactive cognitive model, drawing inspiration from human behavior observed during the image labeling task. Utilizing the VisiTor, the *Vision* module performed visual searches to identify matching objects via template matching, achieving high accuracy and consistently selecting the correct answers. This approach enabled the model to effectively simulate the human cognitive process within the task, incorporating both *Vision* and *Motor* modules along with well-defined declarative and procedural knowledge. However, the model's reliance on simple pattern matching faces challenges with unseen images, as its image labeling capabilities are restricted to data it has previously encountered or specified by modelers, due to the dependence on pre-defined features and sample images for pattern recognition.

Conclusion and Future Work

In this work, we have interpreted human behavior in the image labeling experiment and successfully demonstrated the capability of the interactive cognitive model by integrating the VisiTor (Vision + Motor) framework for performing the image labeling task and predicting completion time. This achievement not only showcases the model's practical applicability but also lays the groundwork for further exploration into simulating human cognitive processes, drawing on observations from the experiment in dynamic labeling task environments. However, certain challenges remain to be addressed in future research.

In future research, it is important to involve a larger number of participants and introduce tasks featuring objects of more complex shapes and a greater variety of classes within the image labeling experiment. A learning curve could be more accurately observed with a significantly increased number of tasks along with a larger pool of participants. Furthermore, to ensure the integrity of individual task assessments, participants will be informed that returning to previous tasks after making a selection will not be permitted. This measure aims to prevent decisions based on direct image comparisons. Additionally, we plan to fine-tune the learning parameters to enhance the retrieval performance of the ACT-R model. Currently, our model skips the learning phase and did not learn the task through the Reference section and Example section. We plan to develop visicon (ACT-R visual scene) and visual objects using vision-language models instead of pre-defining them for the model.

Drawing on the insights from this study aimed at understanding human behavior and developing a cognitive model for image labeling tasks, we take a step forward in our endeavor to replace human labelers with a cognitive model. This transition not only addresses ethical concerns associated with human labeling but also aims to construct a more robust large-scale labeled dataset.

Acknowledgement

the Dorothy Foehr Huck and J. Lloyd Huck Early Career Chair.

References

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341-380.

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press.

Bagherzadeh, A., & Tehranchi, F. (2022). Comparing cognitive, cognitive instance-based, and reinforcement learning models in an interactive task. *Proceedings of ICCM-2022-20th International Conference on Cognitive Modeling.*

Bagherzadehkhorasani, A., & Tehranchi, F. (2023). Automatic Error Model (AEM) for User Interface Design: A new approach to Include Errors and Error Corrections in a Cognitive User Model. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.*

Bothell, D. (2017). Act-r 7 reference manual. Available at act-r psy cmu edu/wordpress/wpcontent/themes/ACT-R/actr7/reference-manual pdf, Accessed February.

Byrne, M. D. (2001). ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Interaction*, *55*(1), 41-84.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition.*

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, *30*(1), 99-120.

Haliburton, L., Ghebremedhin, S., Welsch, R., Schmidt, A., & Mayer, S. (2023). Investigating Labeler Bias in Face Annotation for Machine Learning. *arXiv preprint arXiv:2301.09902.*

Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. *arXiv preprint arXiv:1610.08602.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Laird, J. E. (2019). *The Soar cognitive architecture*. MIT press.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.

Pinker, S. (2003). *How the mind works*. Penguin UK.

Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk.*

Ritter, F. E., Tehranchi, F., & Oury, J. D. (2018). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1488.

Sylolypavan, A., Sleeman, D., Wu, H., & Sim, M. (2023). The impact of inconsistent human annotations on AI driven clinical decision making. *NPJ Digital Medicine*, *6*(1), 26.

Tehranchi, F., & Ritter, F. E. (2017). An eyes and hands model for cognitive architectures to interact with user interfaces. *MAICS, The 28th Modern Artificial Intelligence and Cognitive Science Conference, Fort Wayne, IN: Purdue University.*

Tehranchi, F., & Ritter, F. E. (2018). Modeling visual search in interactive graphic interfaces: Adding visual pattern matching algorithms to ACT-R. *Proceedings of 16th International Conference on Cognitive Modeling (ICCM 2018)*. University of Wisconsin, Madison, WI.

Predictive Algorithms for Individual Reasoning about Possibilities

Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

P. N. Johnson-Laird (phil@princeton.edu)
Department of Psychology, Princeton University Princeton, NJ 08540, USA
Department of Psychology New York University, 6 Washington Place, New York, NY 10003, USA

Abstract

How do people reason about possibilities in everyday life? Most cognitive scientists, including readers of this article, are likely to believe that they rely on a logic, albeit one beyond the grasp of introspection. There are logics for dealing with possibilities—modal logics, and they are useful in software engineering and other domains. This article describes the mental model theory of possibilities and reports two experiments corroborating its central claim that individuals make inferences in default of knowledge to the contrary—a principle inconsistent with all standard modal logics. It also shows that the theory's implementation in a computer program, mModal, accounts for differences from one individual to another in how they reason about possibilities.

Keywords: Modal logic; possibilities; individual differences; mental models, reasoning strategies

Introduction

Consider the following inference:

> It is possible that Ann is in Rio and it's possible that Ben is in Ulm.
>
> ∴ It is possible that Ann is in Rio and that Ben is in Ulm.

This inference condenses two possibilities into one. How reasoners make such an inference about possibilities is a major puzzle. They might rely on some sort of logic for possibilities—a modal logic (e.g., Osherson, 1976). Yet, despite the plausibility of the preceding inference, it is invalid in all standard modal logics (e.g., Chellas, 1980). In this article, we present a theory of modal reasoning based on mental models. It predicts that naive individuals—those who know little or nothing about logic—will accept the inference. Previous studies have borne out other predictions of this theory (see, e.g., Johnson-Laird et al., 2024; Johnson-Laird & Ragni, 2019; Ragni & Johnson-Laird, 2020). The aim of our article is therefore to test the theory's predictions for inferences akin to the example above, and to show that its computer implementation (the mModal program) yields the first working simulation of how individuals differ from one another in their modal reasoning. It does so with just two free parameters.

A major goal for cognitive scientists is indeed to determine how naive individuals make modal inferences. They have carried out empirical studies (e.g., Piéraut-Le Bonniec, 1980), and they have studied how children develop notions of possibility (Shtulman & Carey, 2007) and how adults deduce conclusions about possibilities from factual claims

(e.g., Bucciarelli & Johnson-Laird, 2005; Hinterecker et al., 2016). A pioneering investigation showed that a close relation failed to occur between human reasoning and a subset of a modal logic (Osherson, 1976). What complicates such studies are different sorts of modalities, such as alethic possibilities for inferential relations, epistemic possibilities arising from knowledge, and deontic possibilities based on moral norms and other conventions. A further complication is the number of modal logics. There is a denumerable infinity of them (Chellas, 1980; Ragni & Johnson-Laird, 2018). So, the problem of pinning down which of them, if any, underlies reasoning in daily life seems almost insuperable. It may explain the dearth of pertinent studies.

The argument of the present paper is straightforward. It begins with an account of modal logics, and then with the mental model theory and its implementation in the mModal program. It describes two crucial experiments contrasting modal logic and mental models. And it shows how the program fits the reasoning results of different individuals. Finally, it discusses the implications of these studies for the nature of human cognition and for theories of reasoning.

Modal logics

Modal logics formalize various modalities, which include *possibility* and *necessity* (e.g., Chellas, 1980; Fitting & Mendelsohn, 2023). These logics are useful in software engineering, artificial intelligence (e.g., Kontchakov et al., 2010), and philosophy (e.g., Gödel's ontological proof for God's existence, see Benzmüller & Woltzenlogel Paleo, 2014). Our concern is with how individuals make immediate inferences from a single modal premise to a modal conclusion, and we examined logics based on a system underlying infinitely many others. It is known as system K in honor of the logician Kripke for reasons we explain below. System K and its cognates combine the sentential calculus, which concerns negation (*not*), and compound assertions, such as conjunction (∧), disjunction (∨), and conditionals (→), with modal operators for possibly (◇) and necessity (□). The two modal operators are interdefinable: if something is possible, then it is not necessarily not the case. System K has only a single axiom, which asserts in effect that the necessity of *if A then B* implies that *if A is necessary then B is necessary*:

$$\Box (A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B).$$

The logics built on K introduce further axioms. All these modal logics can rely on a *possible worlds* semantics, and so the assertion:

Possibly A: ($\Diamond$A)

is true provided that A is true in at least one relevant possible world, and the assertion:

Necessarily A: ($\Box$A)

is true provided that A is true in all relevant possible words. Kripke (1963) showed that different assumptions about the relevance of worlds—or their "accessibility" as it is known—parallel the different axioms for modal logics. This discovery led to a burgeoning of modal logics. We say no more about possible worlds, because they seem too vast and too numerous to be a plausible psychological basis for meanings (but cf. the 'miniworlds' of Kripke, 1980, p. 15-20). Moreover, standard modal logics differ from the mental model theory. It postulates that individuals should accept the condensation of consistent possibilities, as in our opening example:

$$\Diamond A \wedge \Diamond B$$
$$\therefore \Diamond(A \wedge B)$$

Yet, such inferences are invalid in all standard modal logics, which admit a counterexample in which the premise is true, but the conclusion is false, i.e., there is a possible world in which A is true, and another possible world in which B is true, but no possible world in which both are true.

The model theory of modal reasoning

The theory is based on an update of ideas going back to 19th century physics and to Craik (1943): humans construct mental models of the world. What counted as a model for Craik was that it computes the same results as reality, and it needs to have no other resemblance to what it models. In contrast, the present theory treats mental models insofar as possible as having the same structure as what they represent: models are *iconic*. One other principle is due to the late Peter Wason (e.g., Johnson-Laird & Wason, 1970), and has been part of the theory in one guise or another from its beginnings (e.g., Johnson-Laird, 1983, see also Kahneman, 2011). There are two systems of human reasoning. One is intuitive (system 1), and the other is deliberative (system 2). The intuitive system has no access to a working memory for intermediate results, and so its computational power is equivalent to a finite-state automaton (Hopcroft & Ullman, 1979). As a result, it tends to use just one *intuitive* model of a possibility at a time. These models represent only those constituents of sentences that are true in the possibility. For instance, an exclusive disjunction, such as:

Either there is a circle or a triangle, but not both.

has a conjunction of intuitive models of possibilities:

○ (Possibility 1)

△ (Possibility 2)

where each row in this diagram represents a different possibility, and each of them holds in default of knowledge to the contrary (Johnson-Laird & Ragni, 2019). The theory therefore predicts that individuals should infer from the disjunction above, e.g.:

It is possible that there is a circle.

System 2, which is elicited when individuals deliberate, has access to human working memory. Hence, it constructs models that represent both what is true and what is false in each model of a possibility. The deliberative models of the disjunction above are as follows, where "¬" is a symbol for negation:

○ ¬△

¬○ △

Negation cannot have an iconic representation, but it is linked to a semantic procedure. These models make clear that the inference of the possibility that there is a circle holds only in case there is not a triangle.

Epistemic possibilities are akin to subjective probabilities: They come in degrees, e.g., an event can be barely possible, highly possible, and at the ends of the scale it is either impossible or certain (see Lassiter, 2017). Unlike standard modal logics, the model theory postulates that an assertion, such as:

It may be raining.

presupposes that it may not be raining. If this presupposition is false, then rain is certain. Table 1 summarizes the mental models of the various sorts of modal assertion.

Table 1: The intuitive (system 1) and deliberative models (system 2) of the four basic epistemic modal assertions about possibilities, where '. . .' denotes an intuitive model with no explicit content, and '¬' denotes negation.

Assertions	Intuitive models	Deliberative models
Possible that A	A	A
	…	¬ A
Possible that not A	¬ A	¬ A
	…	A
Not possible that A	¬ A	¬ A
Not possible that not A	A	A

Because both intuitive and deliberative models hold in default of knowledge to the contrary, the theory predicts three sorts of default inference that reasoners should accept:

1. *It is possible that A or it is possible that B, or both.*
   ∴ *It is possible that A.*
2. *If A then B.*
   It is possible that A.
   ∴ *It is possible that B.*
3. *It is possible that A and it is possible that B.*
   ∴ *It is possible that A and that B.*

None of these inferences is valid in modal logic, but they have been supported in previous studies (e.g., Ragni & Johnson-Laird, 2020). Our present experiments examine the third of these predictions, the condensation of possibilities, both to replicate the phenomenon and to determine whether the mModal program can account for the differences in individual reasoning strategies.

The mModal program for modal inferences

With the help of David Gürth, we developed a computer program mModal, that implements the model theory for modal sentential reasoning. Its source code in Python and relevant data are available online[1]. It adds modal operators, such as *possible* and *certain*, to an earlier program for sentential reasoning, and it deals with alethic, deontic, and epistemic modalities. Among a variety of inferential tasks, which include drawing its own conclusions, it also evaluates given inferences. It allows that reasoners may rely only on system 1 or that in addition they may invoke system 2. In system 1 the intuitive model (cf. Table 1) is built. In System 2 deliberative models (cf. Table 1) are built by the processes described in Figure 2. For an inference, such as:

> *It is raining.*
> ∴ *It is raining or it is hot, or both.*

its implementations in both systems 1 and 2 yield the evaluation:

> ∴ *The conclusion is possible given the premises.*

In contrast, the inference is valid in any normal modal logic and sentential calculus. Both systems in the model theory likewise cope with condensations of possibilities, such as:

> *It is possible that it is raining and it is possible that it is hot.*
> ∴ *It is possible that it is raining and that it is hot.*

System 1 condenses the possibilities in the premise into a single model:

> raining   hot
>   . . .

where the ellipsis represents the other possibilities. So, the conclusion above follows as necessary. If system 2 is called on to evaluate the inference, it constructs explicit models of all four possibilities:

> raining   hot
> raining   ¬ hot
> ¬ raining   hot
> ¬ raining   ¬ hot

So, the correct evaluation is that the inference follows only as a possibility. In other words, system 2 gives a normative account of everyday modal reasoning.

The program implementing the theory allows that knowledge modulates the interpretation of assertions, and one of its effects is to block the construction of a model of a possibility in system 2 (for evidence, see, e.g., Quelhas et al., 2019). Its knowledge-base contains fully explicit models representing, for example, that if and only if a person is married then the person is not single:
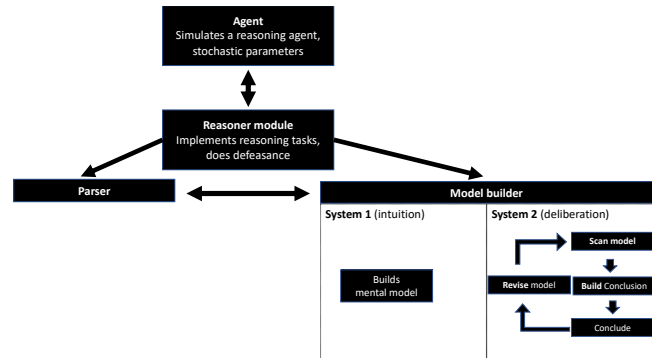
> Person: married ¬ single
>   ¬ married   single

The interpretation of an assertion such as:

> *It is possible that he is married and it is possible that he is single.*

triggers these explicit models. They block the condensation

Figure 1: A diagram of the mModal program implementing the mental model theory. The boxes denote its main components and the arrows denote the flow of control from one component to another.

of the possibilities into one. The model of the possibility that a person is married and single:

> Person: married   single

is not consistent with either of the two explicit models in the knowledge base, and the inconsistency would yield the null model. Hence, the interpretation of the sentence does not condense the two possibilities.

Standard logics evaluate inferences as valid or invalid, where a customary definition is "A valid inference is one whose conclusion is true in every case in which all its premises are true." (Jeffrey, 1981). It allows that any conclusion whatsoever follows as a valid inference from a contradiction—the validity is vacuous, because there are no cases in which the premises are true. Individuals in daily life do not infer arbitrary conclusions from contradictions, and so the logical conception of validity is inappropriate for everyday reasoning.

The model theory and mModal use alethic modals to evaluate inferences. If the premises and a conclusion have no atomic propositions in common, then the program evaluates the conclusion as irrelevant to the premises. Otherwise, a conclusion that describes only one or more possibilities that the premises refer to, and that does not deny any of them, is necessary. A conclusion that is inconsistent with the premises is impossible. And any other sort of conclusion is possible—one that refers to many of the premise models is highly possible (probable). An inference of the sort:

> *The treat is in the cup or the mug. It is not in the cup.*
> ∴ *It is in the mug.*

is necessary and logically valid. In contrast, this sort of inference:

> *The treat is in the mug.*
> ∴ *It is in the mug or the world will end shortly, or both.*

is not necessary, though it is logically valid. People reject it. Alethic necessity can be problematic in certain cases. An inference such as:

---

[1]https://github.com/CognitiveComputationLab/cogmods/tree/master/modal

∴ *It is raining or else it's cold, but not both.*

The intuitive models of the premise correspond to the conjunction of three default possibilities:

    raining

                cold

    raining   cold

The conclusion holds in first two of these models, but not in the third model—indeed, it denies this possibility. So, strictly speaking, it is not a necessary inference, but we refer to it as a case of *weak* necessity because it fails only in denying a possibility that the premise allows. The converse inference:

    It is raining or else it's cold, but not both.

    ∴ *It is raining or it's cold, or both.*

is logically valid, but it is not necessary, because nothing in the premise describes the possibility that both events occur.

Strategies in modal reasoning

The theory implies that people should differ in the conclusions that they draw, and so for the first time we examine individual differences in modal reasoning. To understand our procedure, we need to explain the overall organization of the mModal program. It has four main modules: an agent, a reasoner, a parser, and a model builder. Figure 1 depicts these modules, describes their main functions, and shows the relations between them.

The agent uses two free parameters to model reasoning. One parameter (*sigma*) is the probability that system 2 is invoked to make an inference, and so its values range from 0 to 1. The other parameter (*gamma*) is the probability that the program uses weak necessity to assess inferences. A third but pertinent distinction is whether individuals evaluate conclusions as necessary or as possible: the program allows both options. Table 2 summarizes the six alternative strategies that mModal predicts.

Table 2: Eight alethic reasoning strategies implemented in mModal for the evaluation of inferences in which premises and conclusion share at least one atomic proposition

System	Given the premises, it can evaluate the conclusion as:
Intuitive (1)	necessary
	weakly necessary
	possible
	impossible
Deliberative (2)	necessary
	weakly necessary
	possible
	impossible

In order to compare the theory's predictions about how individuals' differ in their reasoning strategies, and also to examine how well their performance fitted standard modal logics, we used the CCOBRA system (see https://github.com/CognitiveComputationLab/ccobra) to analyze mModal. It examined the program's predictions, manipulating the free parameters in its Agent component to make an optimal fit with the data from each participant, and it computed the numbers of participants at each percentage level in the accuracy of mModal's prediction. It likewise computed the numbers of participants at each percentage level in matching the inferences common to four modal logics (systems K, T, S4, and S5), using the Modal-Logic-Tableaux-Solver2, which Joey Thaidigsman devised. We present these results for each experiment.

Experiment 1: Condensation of modals

The experiment examined a representative set of inferences from a single modal premise to a single modal conclusion to contrast the mental model theory's predictions about condensations with those of standard modal logics. The main set of 18 inferences included cases of this sort:

    Premise: *It is possible that A and it is possible that B.*

    Conclusion: *It is possible that A and that B.*

where A and B are sensible assertions (see below). The theory predicts that individuals should make this condensation in default of knowledge to the contrary. It is invalid in standard modal logics because A and B could be inconsistent with one another, and thus jointly impossible. In the experimental inferences, the modals in the premise and conclusion were always the same, as were the connectives. The modals were of three sorts: *possible*, *necessary*, and *impossible*; and the connectives were of three sorts: *and*, exclusive *or*, and inclusive *or*. There were two versions of these nine sorts of inference. In one version, both clauses in the premise had its own modal and the conclusion had single modal, i.e., a condensation such as:

    ◇A and ◇B ∴ ◇(A and B)

In the other version, the premise and conclusion were swapped round:

    ◇(A and B) ∴ ◇A and ◇B

The theory predicts that participants should accept all 18 of these inferences, whereas in the modal logics eight of them are invalid (see Table 3 below). As a control, the experiment examined a further nine inferences that the theory predicts should be rejected, but here we focus on those that it predicts should be accepted.

The contents of the inferences were sensible everyday assertions about the locations of two individuals in well-known cities, and the participants' task was to answer the question: Does the premise imply that the conclusion is true? They responded either 'yes' or 'no'. The inferences were presented to each of them in a different random order. The 67 native English speakers in the experiment had no knowledge of logic. They were tested on Amazon's Mechanical Turk.

Results

Table 3 presents the results for the inferences for which the mental model theory predicts acceptances. The participants made 80% of the theory's predicted evaluations but only 56% evaluations that fit modal logic, and the difference was robust (Wilcoxon test, $z < 6.0$, $p < .00001$). For inferences in which the mental model theory diverged from modal logic, the predicted 'yes' evaluation occurred on 84% of trials, which

Table 3: The percentages of participants accepting the 18 immediate inferences in Experiment 1 ($N = 67$) for which the mental model theory predicts acceptances. Percentages in bold are for inferences that are invalid in standard modal logics (systems K, T, S4, and S5).

| | | Inferences: | Modal operators in inferences | | |
		M is a modal	*Possible*	*Necessary*	*Impossible*
Sentential connective in the premise	Conjunction	MA∧MB ∴M(A∧B)	**92**	88	83
		M(A∧B) ∴MA∧MA	94	88	**87**
	Inclusive *or* disjunction	MA∨MB ∴M(A∨B)	90	86	**85**
		M(A∨B) ∴MA∨MA	87	**85**	87
	Exclusive *or* disjunction	MA∨̄MB ∴M(A∨̄B)	90	**85**	87
		M(A∨̄B) ∴MA∨̄MA	**81**	**67**	**83**

was reliably greater than logical 'no' evaluations (Wilcoxon test, $z = 5.9$, $p < .00001$). The control filler inferences tended to elicit rejections of the inferences (75%).

Figure 2 presents bee-boxes comparing how well mModal and the modal logics accounted for the performance of the different participants. As it shows, mModal's predictions are more accurate about individual participants (median of 69% accuracy) than modal logic (median of 52% accuracy). The difference was reliable (Wilcoxon test based on participants, $z = 7.85$, $p < .1^{15}$, $r = .96$).

We compared the fit of the identified strategies to each participants response pattern: It showed that the majority of them relied on system 1: 52% used it to evaluate necessary conclusions, 24% used it to evaluate possible conclusions, and 10% used it to evaluate weakly necessary conclusions. The remaining 14% used system 2 and were divided in almost equal proportions among the three different modal evaluations.
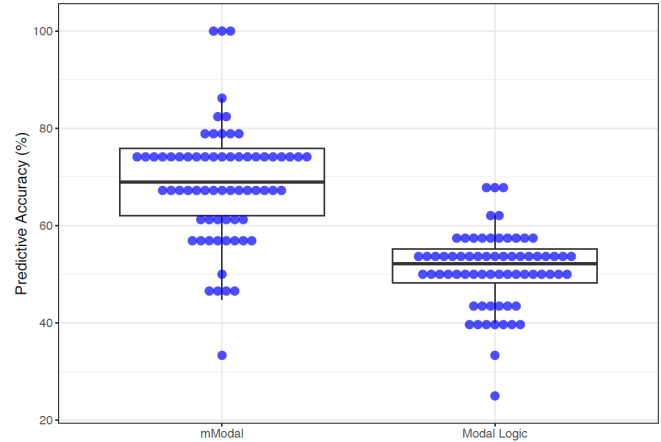
Figure 2: The accuracy of mModal and modal logic in predicting the evaluations of individual participants in Experiment 1 ($N = 67$): each dot shows the proportion of a participant's evaluations that mModal predicts (the left box) and that the modal logics endorse (the right box). Each bee-box shows the four quartiles of the participants, with the two central quartiles divided by a horizontal line denoting the median, and the two whiskers denoting the upper and lower quartiles.

Experiment 2: The consistency of condensations

This experiment tested the mental model theory's principle that individuals condense possibilities only when they are consistent. The experiment compared assertions that were consistent with those that were inconsistent and that should block the inference, e.g.:

Premise: *It is possible that Tom is single and it is possible that Tom is married.*

Conclusion: *It is possible that Tom is single and married. Does the premise imply that the conclusion is true?*

We tested 53 participants from the same population as before. They acted as their own controls and evaluated 12 condensations: inferences from clauses in the premises that were each asserted to be possible to a conclusion that was asserted to be possible. Half the inferences had conclusions that were consistent, e.g.: *It is possible that Tom is married and that Ben is single;* and half had conclusions that were inconsistent, e.g., *It is possible that Tom is married and that he is single.* Within each of these halves, the inferences were based either on conjunctions in both the premise and conclusion, or on inclusive disjunctions in both of them. The problems were presented in a different random order to each participant. The procedure was identical to the one in the previous experiment.

Results

Table 4 presents the results of the experiment. They corroborated mModal's predictions: The participants accepted the conclusions for 74% of consistent conclusions,

but for only 21% of inconsistent conclusions (Wilcoxon test, $z = 5.7$, $p < .00001$).

Table 4: The percentages of the participants' acceptances ('yes' evaluations) of the four sorts of inference in Experiment 2 ($N = 53$) depending on whether the connective, in the premises and conclusion, was a conjunction or an inclusive-or, and on whether the clauses A and B, were consistent with one another or inconsistent.

	Status of the conclusions	
Sort of inference	Consistent	Inconsistent
◇A *and* ◇B ∴◇(A *and* B)	89	27
◇A *or* ◇B ∴◇(A *or* B)	58	15

The participants also tended to accept inferences based on conjunctions more often than those based on disjunctions (Wilcoxon test, $z = 3.68$, $V = 108$, $p < .0002$, $r = 0.36$)—a known phenomenon that the mental model theory predicts (García-Madruga et al., 2001). The interaction between these two variables was also significant with conjunctions yielding a larger difference between consistent and inconsistent inferences than disjunctions (Wilcoxon test, $z = 3.43$, $V = 81.0$, $p < .0001$, $r = 0.34$). We surmise that the greater difficulty of disjunctive inferences led to a 'floor' effect on accuracy with inconsistent inferences.

Figure 3 shows the accuracy of mModal and modal logic using the same sort of bee-plots as before. As they indicate, mModal makes a greater proportion of accurate predictions (median of 72%) than modal logic (median 50%; Wilcoxon test $z = 6.8$, $p < .10^{-12}$, $r = .84$). The automated analysis showed that all the participants tended to evaluate the necessity of inferences, with 61% of them relying on system 1, and the remainder relying on system 2.
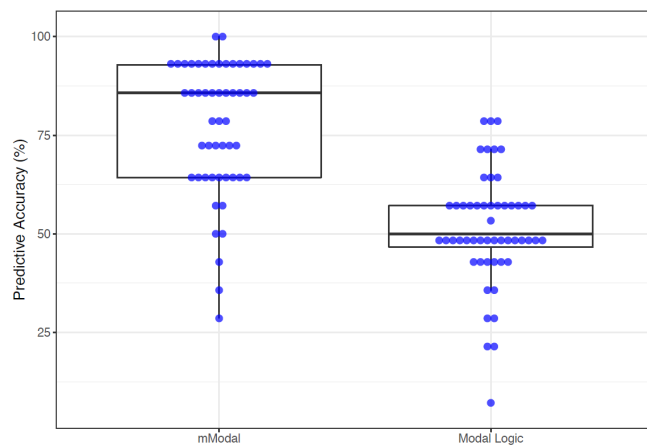

Figure 3: The accuracy of mModal's predictions and modal logic's predictions: each dot shows the proportion of a participant's evaluations that the relevant approach predicted correctly in Experiment 2 ($N = 53$).

General Discussion

The experimental evidence corroborates three aspects of the mental model theory of modal reasoning, and its computer implementation, mModal. Contrary to standard modal logics such as system K, reasoners tend to condense possibilities into one (Experiment 1), but they do not do so when inconsistent predicates would hold for the same individual (Experiment 2), and their inferential strategies usually rely on the intuitive system for reasoning, and on evaluating whether a putative conclusion is necessary (Experiments 1 and 2). So, they accepted inferences such as:

> It is possible that Ben is alive and it is possible that Tom is dead.
> ∴ It is possible that Ben is alive and that Tom is dead.

But, they rejected condensations if Tom's respective possibilities are being *alive* and being *dead*. Earlier results showed that individuals accept inference of the sort:

> It is possible that A or it is possible that B, or both.
> ∴ It is possible that A.

and:

> If A then B.
> It is possible that A.
> ∴ It is possible that B.

Hence, the evidence bears out the theory of mental models, and rules out standard modal logics as playing even a normative role in everyday reasoning. People have a natural tendency to treat epistemic possibilities as akin to subjective probabilities (see also Lassiter, 2017), as presupposing the possibility of their denials, and as holding only in default of knowledge to the contrary.

Despite its predictive success, the mModal program has yet to embody several aspects of the mental model theory, such as its account of counterfactual assertions. A major lacuna in the theory itself is its rudimentary treatment of multiple modal operators in the same assertion, e.g.:

> It's possible that the conclusion may follow of necessity.

Likewise, the theory makes many predictions that have yet to be tested. But no other theory of reasoning accounts for the present results. Theories that have replaced logic with the probability calculus (see, e.g., Oaksford & Chater, 2020) have yet to address modal reasoning, and probability cannot account for the meaning of permissibility, i.e., deontic possibility. Meanwhile, it seems safe to conclude that normal modal logics yield implausible accounts of naive reasoning from premises containing modal operators.

Acknowledgments

References

Benzmüller, C., & Woltzenlogel Paleo, B. (2014). Automating Gödel's ontological proof of God's existence with higher-order automated theorem provers. In T. Schaub, G. Friedrich, & B. O'Sullivan (Eds.), *ECAI 2014, Frontiers in Artificial Intelligence and Applications* (Vol. 263, pp. 93-98). IOS Press.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology, 50*(2), 159-193.

Chellas, B. F. (1980). *Modal logic: an introduction.* Cambridge university press.

Craik, K. (1943). *The Nature of Explanation.* Cambridge University Press.

Fitting, M., & Mendelsohn, R. L. (2023). *First-order modal logic* (2nd ed.). Springer.

García-Madruga, J. A., Moreno, S., Carriedo, N., Gutiérrez, F., & Johnson-Laird, P. N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *The Quarterly Journal of Experimental Psychology Section A, 54*(2), 613-632.

Hinterecker, T., Knauff, M., & Johnson-Laird, P. N. (2016). Modality, probability, and mental models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1606.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation.* Addison-Wesley.

Jeffrey, R. C. (1981). *Formal logic: Its scope and limits* (2nd ed.). McGraw-Hill.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Johnson-Laird, P. N., Byrne, R. M. J., & Khemlani, S. (2024). Models of possibilities instead of logic as the basis of human reasoning. *Minds and Machines*, *34*, 19.

Johnson-Laird, P. N., & Ragni, M. (2019). Possibilities as the foundation of reasoning. *Cognition, 193*, 103950.

Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive psychology, 1*(2), 134-148.

Kahneman, D. (2011). Thinking fast and slow. Farrar, Strauss, & Giroux.

Kontchakov, R., Wolter, F., & Zakharyaschev, M. (2010). Logic-based ontology comparison and module extraction, with an application to DL-Lite. *Artificial Intelligence, 174*(15), 1093-1141.

Kripke, S. A. (1963). Semantic considerations on modal logic. *Acta philosophica fennica*, *16*, 83-94.

Kripke, S. A. (1980). *Naming and necessity*, 2nd Ed. Cambridge, MA: Harvard University Press.

Lassiter, D. (2017). *Graded modality: Qualitative and quantitative perspectives*. Oxford University Press.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual review of psychology, 71*, 305-330.

Osherson, D. N. (1976). *Logical Abilities in Children*. (Vol. 4). Erlbaum.

Piéraut-Le Bonniec, G. (1980). *The development of modal reasoning: Genesis of necessity and possibility notions*. Academic Press.

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2019). The analytic truth and falsity of disjunctions. *Cognitive Science, 43*(9), e12739.

Ragni, M., & Johnson-Laird, P. (2018). Reasoning about possibilities: human reasoning violates all normal modal logics. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 2309-2314). Cognitive Science Society.

Ragni, M., & Johnson-Laird, P. N. (2020). Reasoning about epistemic possibilities. *Acta Psychologica, 208*, 103081.

Shtulman, A., & Carey, S. (2007). Improbable or impossible? How children reason about the possibility of extraordinary events. *Child development, 78*(3), 1015-1032.

Towards a Comprehensive Summary of Senses for Cognitive Architectures

Frank E. Ritter (fer2@psu.edu)
Serhii Serdiuk (sjs8152@psu.edu)
College of IST, Penn State
University Park, PA 16802 USA

Abstract

It is widely accepted that there are five senses. There, however, appear to be many more. This paper provides a comprehensive list of sensors that will be found in a complete cognitive architecture. We also briefly note how widely used these senses have been and which ones could yet be implemented in an architecture.

Keywords: senses, cognitive architecture

Introduction

It is widely accepted that there are five senses (e.g., Sekuler & Blake, 2001). There has always seemed to be more. A recent review of cognitive architectures (Kotseruba & Tsotsos, 2020; in press) provides a wide enough review to suggest that there are other senses (ways of knowing the state of the world or the agent using a sensor) that have sometimes been included in cognitive architectures. Upon further reflection, there appear to be far more than five senses. This paper describes a comprehensive list of sensors that might be found in a complete cognitive architecture. We also briefly note how widely used these senses have been.

There are also numerous measures of the body that are not available or not completely available to cognition, such as calorie needs, water intake, blood sugar, and insulin levels. These measures have to be measured externally, suggesting that not all aspects of the human state are available directly as a sense.

Thus, the first column of Table 1 provides the start of a comprehensive list of human senses for inclusion in a comprehensive cognitive architecture. The table is divided into external senses, which measure the world, and internal senses that tell cognition about the body. The second column of Table 1 represents corresponding human sensory systems.

Additionally, aspects such as perception of time, emotional senses, and social tension perception could be considered integral aspects of human cognition that are not traditionally categorized as the five basic senses. These components may draw from multiple senses and aspects of cognition, contributing to the complex interplay of cognitive processes and should be acknowledged in discussions of cognitive architectures.

There are some further senses that some animals have, such as a measure of magnetic fields. If we are to build a cognitive architecture for animals, we will need those senses. If we are building a superhuman model, for example, or a robot, we may wish to have those as well. Or we way may find that humans have vestigial versions of these senses.

It is worth noting that the information presented in Table 1 is not exhaustive, and our scientific understanding is constantly evolving. Recent research suggests the possibility of a subconscious magnetic sensory system in humans (Wang et al., 2019), challenging previous assumptions.

The remainder of this paper explains each of these senses in more detail. With the description in hand, we then discuss the implications for modeling cognition and performance.

This is a conference paper, so the review will necessarily be brief and preliminary. If we have left out major results, we beg readers to allow us some grace and provide feedback—to not bite our finger but note where we are pointing.

Brief Description of the Senses

In this section, we briefly describe each sense in Table 1 and how they have been used in cognitive architectures. When we do not note a documented sense, we do not know of work using this sense in generative cognitive modelling.

External senses

These senses tie the body to world in various ways.

1 Sight. It is expected that the visual modality takes first place here, because it provides the fastest way to transmit information about objects in the environment. This modality is also most widely used in human-machine interfaces (HMI) (e.g., Rydström & Bengtsson, 2007). The importance of vision for obtaining information about the world around us and surviving is evidenced by the fact that the brain devotes more space and resources to processing information coming from the visual system than to information coming from all other senses combined. The visual modality in cognitive architectures is most often implemented using physical sensors or simulations (Kotseruba & Tsotsos, 2020).

2 Hearing. Audition is the second most frequently used modality in modern HMIs. It is often used to provide an adjunct to the visual channel, to unload the visual channel for receiving information and duplicating it. Auditory modality is recommended for use in cases of difficulty or unavailability of the visual system like high positive accelerations, oxygen deprivation, unnatural lighting or its absence, the need to change working position, etc.

There are two ways to transmit auditory information to a person: verbal and audio signals. Words are a fundamental aspect of auditory communication and can greatly impact human cognition and behavior. When presented verbally, words engage cognitive processes related to language comprehension, semantic interpretation, and working memory.

Table 1. A relatively complete list of human senses.

Senses	Sensory systems
External/Distal	
1 Sight	Visual system
2 Hearing	Auditory system
3 Taste	Gustatory system
4 Smell	Olfactory system
5, 6, 7, 8 Touch	Somatosensory System
- touch	- mechanoreceptors
- temperature	- thermoreceptors
- nociception (pain, skin)	- nociceptor
- vibrations	- mechanoreceptors
Internal	
9 Sense of balance or equilibrioception, vertigo (loss of balance)	Vestibular System
10 Air pressure (popping ears)	Auditory system Vestibular System
11 Time (passage of)	Interoceptive system and other systems
12 Proprioception (body position, all limbs & head)	Somatosensory System - proprioceptors
13 Energy level/fatigue	Interoceptive system and other systems
14 Need for sleep	Interoceptive system and other systems
15 Emotions (e.g., I am feeling anger)	Interoceptive system and other systems
16 Body temperature	Interoceptive system - thermoreceptors
17 Nociception (pain, inside)	Interoceptive system, Somatosensory System - nociceptors
18 Need for air up to suffocation	Interoceptive system - chemoreceptors
19 Thirst	Interoceptive system - osmoreceptors and other systems
20 Hunger	see Fig. 2
21 Nausea (Stomach, gut, lower gut)	Interoceptive system - mechanoreceptors - chemoreceptors - osmoreceptors
22 Need to void (urine)	Interoceptive system - mechanoreceptors
23 Need to void (solid)	Interoceptive system
24 Need to void (gas)	- mechanoreceptors - chemoreceptors

Sound signals, including tones, beeps, alarms, and melodies, serve various purposes in HCI and can significantly influence human cognition and behavior. The characteristics of sound signals, such as frequency, amplitude, duration, and temporal pattern, play a crucial role in determining their impact on the user.

The auditory modality in cognitive architectures is most often implemented simulating the results (hear: "Hello") or using physical sensors (Kotseruba & Tsotsos, 2020). ACT-R, for example, has a simple ear (Byrne, 2001). A design for a more complete auditory system has been contemplated (Ritter, Brener, Bolkhovsky, 2023). Most architectures do not yet hear all these kinds of sounds or even simulate them.

3 Taste. At its core, the gustatory system governs the execution of behavioral sequences necessary for locating, inspecting, and ingesting food, a critical function for the survival of all animals.

Taste perception involves complex cognitive processes that go beyond mere sensory input. Additionally, taste can evoke emotional and physiological responses, further shaping human behavior. For instance, the taste of certain foods may elicit pleasure or disgust, leading to corresponding emotional reactions and can influence decision-making processes regarding food selection and consumption. MicroPsi appears to be the only system with taste (Bach, 2008) because it models finding food, discriminating, and eating.

4 Smell. The olfactory system plays a significant role in human life but is underrepresented in scientific research (van Hartevelt & Kringelbach, 2015). This system not only helps us select food and ensure survival, but is also unique in its structure and functioning. Unlike other sensory systems, olfactory information does not pass through the thalamus, but goes directly to cortical areas such as the orbitofrontal cortex.

Numerous studies point to a deep relationship between the olfactory system and emotions. This is due to the fact that they use common brain structures for processing: amygdala, hippocampus, insula, anterior cingulate cortex and orbitofrontal cortex (Soudry et al., 2011). Smells can evoke emotional reactions, positive or negative, and can lead to nausea, which is then sensed and perhaps amplified as an emotional feeling. Some scents can systematically evoke certain emotions, and a person's emotional state can influence the perception of odors in the environment.

Impaired sense of smell can lead to anhedonia. This phenomenon leads to a decrease in motivation and pleasure and is a symptom of many mental illnesses: schizophrenia, Parkinson's disease, eating disorders, borderline personality disorder, etc. (Pelizza & Ferrari, 2012).

Smell is represented as a sensory modality in the DAC (Mathews et al., 2009), GLAIR (Shapiro & Kandefer, 2005) and PRS (Taylor and Padgham 1996) architectures. PSI (Dörner, 2000) and MicroPsi (Bach, 2008) also has it because their models find food and eat.

5,6,7,&8 Touch+Temperature+Nociception+Vibration. A human receives the majority of information in control systems through visual and auditory analyzers. The tactile analyzer is relatively rarely used in architectures, despite its immense potential. In real life, humans perform numerous gnostic (touch, palpation, contour following, etc.), controlling, and identifying movements with their hands.

There are numerous theories about cutaneous sensitivity, which are largely contradictory. However, it is established that cutaneous receptors (see Somatosensory system below)

include receptors responsible for modalities such as touch, pressure, vibration, temperature, and nociception (pain).

Humans can fairly accurately determine the location of a stimulus, the distance on the skin between points of stimulation, distinguish the degree of stimulation, etc. Thus, with the help of a multidimensional signal (e.g., a combination of vibration frequency+point of stimulation+amplitude+interval between signals), a significant amount of information about the control object can be transmitted (such as its state, position in space, speed, and assessment of time intervals).

The vibrotactile modality is often used in devices such as smartphones and tactile navigation devices. The cutaneous sensory system has also been used for recognizing visual patterns using discrete points, duplicating the visual modality (Lindsay & Norman, 1972). In some cases, tactile perception surpasses vision. Through touch, a human can assess the weight of objects, and determine their temperature and hardness directly.

Many robotic platforms are equipped with sensor bumpers necessary for effective problem-solving in movement and ensuring safety during motion. The touch modality is implemented only in 21% of cognitive architectures (Kotseruba & Tsotsos, 2020).

Internal senses

These are senses that the architecture would use to recognize the state of cognition or of the body.

9 Sense of balance. Just as vision provides information about the external environment, the sense of balance, or equilibrioception, provides internal feedback about the body's orientation and movement in space, joint position, muscle force, and effort. Our sense of balance helps us maintain stability and adjust our posture to prevent falls and maintain equilibrium. Many of us can remember the emotions that arise even when we feel slightly dizzy, especially when we realize that this could happen in dynamic situations, such as driving a car. Even short-term difficulties perceiving the position or movement of body parts can lead to difficulty performing tasks that require precise movements or spatial navigation. The vestibular system is an important component of proprioception and is responsible for maintaining static, mixed or dynamic balance. A human can improve balance and movement perception by training proprioception (Zsolt, 2018).

10 Air pressure (popping ears). In addition to external stimuli like sound waves, our bodies also perceive changes in air pressure. Similar to how we perceive changes in external air pressure, such as when flying in an airplane or diving underwater, our internal sensors detect shifts in atmospheric pressure and respond accordingly by "popping" our ears and equalizing this pressure. In cognitive modeling of certain types of activities, knowledge of how to reset the sensor after pressure changes will be useful to match human behavior.

11 Time (passage of). The perception of time is fundamental to human cognition (e.g., Taatgen, Van Rijn, & Anderson, 2007; Stine, Klein, & Yatko, 2001; Wittmann, 2009) as it provides the framework within which events are ordered, episodic memories are formed, and plans are made. Under-

standing how the brain processes and perceives time is crucial for developing accurate cognitive models, especially in areas such as decision-making, planning, and memory.

12 Proprioception (body position, all limbs, and head). This internal sense allows us to coordinate movements and maintain balance without constant attention to external visual or tactile signals. Robots use this sense, but we know of no models that do.

13 Energy level/fatigue. Similarly to how we perceive external stimuli like temperature or texture through touch, our bodies internally perceive changes in energy level and fatigue in ourselves. This internal feedback informs us of our body's physiological state, ranging from feelings of alertness (which may be a separate sense) and vitality to sensations of tiredness and depletion, influencing our physical and mental performance. There are quite a few known models of physical fatigue. Liang et al. (2009) considered, for example, 24 static and three dynamic muscle fatigue models. Taking into account fatigue during mental work or social interaction is no less important, but no models are tied to cognition but for PSI.

Patzelt and Shepherd (2024) recently presented a fatigue model of social venturing and showed that social project fatigue leads to an entrepreneur's disengagement from goals and decreased sensitivity to social issues, diminishing the entrepreneur's prosocial motivation to achieve goals and/or prompting them to abandon social projects altogether. Correct assessment and modeling of various types of fatigue will be necessary to determine the optimal mode of work and rest and will be very important for cognitive modeling long-term tasks.

14 Sleep, need for. Comparable to how we respond to external cues like darkness or noise to initiate sleep, our bodies internally perceive signals indicating the need for rest and sleep. While this sense may be imperfect and not always align with external factors, such as work schedules or environmental conditions, it plays a crucial role in regulating our sleep-wake cycle and overall well-being. The need for sleep is essential for cognitive recovery, memory consolidation, and our overall brain health.

15 Emotions. Emotional senses, including the recognition and interpretation of emotions in oneself and others, play a central role in human social interaction, decision-making, and overall well-being. Some of this processing is just cognitive. This information might appear from vision, but noticing the emotions in oneself might be seen as a sense that varies across people and may be tied to the gut.

Social tension perception (external) refers to the ability to sense and respond to social cues, hierarchies, and dynamics in interpersonal interactions. This aspect of cognition is critical for navigating complex social environments, forming alliances, and predicting others' behavior. Incorporating social tension perception into cognitive models can lead to more realistic simulations of human behavior and societal dynamics. Knowing yourself (and others) is perhaps now a type of sense to put on our list.

Emotions and internal sensations are closely intertwined. Similar to how external stimuli can elicit emotional responses, our internal experiences of emotions such as anger or shame can be a form of sensory perception. Just as we interpret external stimuli to generate emotional responses, our internal emotional states also provide feedback about our psychological well-being and inform our thoughts, behaviors, and decision-making processes.

This highlights the importance of developing emotional models for cognitive architectures as well as in human-robot interactions. An emotion ontology (e.g., www.ebi.ac.uk/ols4/ ontologies/mfoem) can be a starting point for this.

16 Body temperature. Analogous to how we perceive external temperatures, our bodies internally sense changes in our body temperature. This internal feedback informs us of our body's thermal state, whether we feel warm or cold, and triggers physiological responses such as shivering or sweating to maintain homeostasis.

Humans and animals have several internal highly sensitive (molecular) "thermometers", presented in the form of transient receptor potential channels (TRP), which ensure the maintenance of internal body temperature with minimal energy expenditure.

Temperature sensitivity has not yet been sufficiently studied. For example, "mild cooling is detected by the menthol-sensitive TRPM8 ion channel, but how painful cold is detected remains unclear" (Buijs & McNaughton, 2020).

17 Nociception (pain). A human has pain-sensitive receptors also in the internal organs. Experts distinguish between the concepts of "nociception" and "pain" (Sneddon, 2018).

Nociception is the ability to detect and respond to various stimuli. Our body instantly responds to the stimulus with protective reflexes, which are called nocifensive. If these reactions become long-lasting and change our behavior, it may be a sign of discomfort associated with pain. Thus, pain is not only a physical sensation, but also a complex emotional and behavioral experience. Here, we count internal pain.

Additionally, there is an understudied role for passive nociception. Passive nociception involves the participation of inactive nociceptors in controlling our behavior. They seem to "push" us and direct us to ensure that actions do not cause pain or injury (Armstrong, 2024). This, for example, explains periodic stretching during prolonged sitting. There may be multiple types of receptors. Autonomic systems may provide an analogy.

18 Need for air. Comparable to how we respond to external stimuli like smoke or carbon monoxide by seeking fresh air, our bodies internally perceive the need for oxygen and respond to insufficient oxygen levels with sensations of suffocation or breathlessness. This internal sense of respiratory distress prompts behaviors to ensure adequate oxygen intake, such as adjusting breathing patterns or seeking oxygen-rich environments. This is an important sense and can be confused by carbon monoxide and nitrogen.

19 Thirst. Analogous to how we respond to external cues like dryness or saltiness by feeling thirsty, our bodies internally sense the need for hydration through the sensation of thirst.

Osmoreceptors of the interoceptive system play an important role in detecting water imbalance and drinking behavior in humans. This internal feedback signals dehydration and prompts behaviors to seek and consume fluids to maintain fluid balance and prevent dehydration. Dehydration can lead to cognitive deficits, including problems with concentration, memory, mood regulation, etc.

Mechanoreceptors of the interosceptic system, which respond to changes in pressure in the hollow organs, help us perform reflex acts of urination and defecation. ACT-R/Phi has worked in this area.

20 Hunger. See example description below.

21 Nausea (Stomach, gut, lower gut). A major aspect of human thought and knowledge has been left out of all cognitive architectures, even PSI (Bach, 2008; Dörner & Güss, 2013). This aspect is the gut brain (Mayer, Nance, & Chen, 2022) that mobilizes the movement of food through the body with around 100 million nerve cells, and has concurrent effects on mood and that informs or indirectly leads to changes in multiple internal measures, such as need to void.

Nausea serves as a warning sign of digestive disturbances or toxin exposure, influencing cognitive and behavioral responses. Nausea and hunger can impair concentration and decision-making, leading to decreased performance and motivation. Osmoreceptors, together with chemoreceptors, help a human determine the presence and concentration of harmful substances in the body. Mechanoreceptors are responsible for reflex coughing and vomiting.

22 Need to void (urine). The urge to urinate signals the body's need to eliminate liquid waste products and maintain urinary function. Ignoring or delaying this sensation can lead to discomfort and distraction, affecting cognitive focus and productivity, and has even killed people (e.g., Tycho Brahe). Acknowledging and responding to the need to void is essential for maintaining physical comfort and supporting cognitive well-being.

23, 24 Need to void (solid, gas). Humans, like animals, can generally know that they need to void their solid waste. They also often, but not infallibly know whether it will be solid, liquid, or gas.

Human Perceptual Systems

To understand how a human interacts with the world and how we can model this process, we briefly consider a few sensory systems. We leave out the most commonly covered, vision and hearing. This survey, in turn, will help us create a more advanced and comprehensively meaningful cognitive architecture.

Somatosensory System

Purpose: Provides the brain with information about various sensations inside and around the body, including touch, pressure, vibration, pain, temperature, and body position in space. It includes the cutaneous sensory system and the musculoskeletal sensory system.

Main Function: Transmitting sensations for awareness and response to external and internal stimuli.

Receptors: Mechanoreceptors such as Merkel discs and Meissner's corpuscles (sense touch, pressure and vibration), thermoreceptors (sense temperature), nociceptors (sense pain), proprioceptors (perceive body position in space).

As mentioned above, proprioception ranks third in number of implementations after vision and symbolic input. Therefore, we will present a brief look at the proprioceptor sensory system as a part of Somatosensory System:

Purpose: Provides the brain with information about movement, the position of body parts relative to each other, and the force necessary to perform movements.

Main function: Movement planning and control.

Receptors: Proprioceptors (this is a type of mechano-receptor) - in muscles, joints, tendons, ligaments and connective tissues.

Gustatory system

Purpose: Provides the brain with information about the taste of food.

Main Function: Perception of taste qualities of food.

Receptors: Proprioceptors in the tongue and other areas of the oral cavity react to various taste qualities of food, including sweet, salty, sour, bitter, and umami.

Olfactory system

Purpose: responsible for perceiving odors.

Main Function: Assessing appetitive aspects of food. The Olfactory System plays a vital role in evaluating the aromatic properties of food and can influence the desire to consume particular items.

Receptors: Olfactory receptors located in the nasal cavity. These receptors respond to chemical substances in the air we inhale. They detect various aromatic molecules.

Vestibular System

Purpose: responsible for perceiving body position and movement in space. It includes the ear and a set of neural connections that help maintain balance, coordinate movements, and control spatial orientation.

Main Function: The main function of this system is to maintain the body's balance and orientation in space. It allows an individual to assess whether they are in a vertical position, moving, or stationary.

Receptors: Vestibular receptors located in the inner ear. These receptors respond to head movements and body posture. They perceive acceleration and changes in head position, enabling the assessment of the body's position in space.

Interoceptive System

Purpose: Provides the brain with information about internal physiological states of the body, such as hunger, thirst, fatigue, pain, temperature, organ conditions, and other biological aspects. This system plays a key role in self-awareness of physiological states and responding to them.

Main Function: Perception of internal sensations.

Receptors: Interoceptive receptors in various organs including the stomach, intestines, and other internal parts, detecting glucose levels, pressure, temperature, and other internal parameters.

A semantic network (Fig. 1) was developed to formally describe this knowledge. It describes how a Human uses sensory systems to perceive the outside world, including: the Visual, Auditory, Gustatory, Somatosensory, Olfactory, Vestibular, and Interoceptive systems.
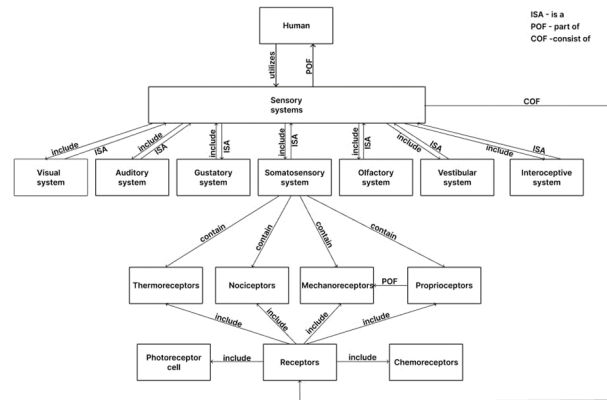
Figure 1. Formal representation of knowledge about the human sensory system.

In addition, the Sensory System consists of the following set of receptors: chemoreceptors, mechanoreceptors (that in turn include proprioceptors), photoreceptors, thermoreceptors, nociceptors. Figure 1 shows only a fragment of the semantic network to not make the picture noisy, only Somatosensory system receptors are indicated.

How do humans use their sensory systems?

Let's consider this using the example of a feeling that each of us has experienced at least once in our lives—the feeling of hunger. The main human sensory systems associated with feelings of hunger and satiety are:

• Taste. It is associated with the perception of the taste of food. Receptors in the tongue and other parts of the oral cavity respond to different taste qualities of food

• Olfactory system. The sense of smell plays a key role in assessing the appetizing aspect of food. Smells can greatly influence appetite and anticipation of food intake

• Somatosensory system. It is responsible not only for sensations associated with the bodily senses, but also for sensations associated with digestion. Mechanoreceptors in the stomach and intestines respond to organ distension and chemical changes associated with food intake. Chemoreceptors found in the stomach, intestines and other parts of the digestive tract respond to chemical changes in the body, including changes in the levels of hormones such as ghrelin (the hunger hormone) and leptin (the satiety hormone). They help regulate feelings of hunger and satiety by interacting with the hormonal systems, which are also involved in this regulation.

Based on information received from sensory systems, the brain decides how the body should respond. These reactions may include:

• Appetite regulation: The brain can regulate appetite levels by influencing hormonal systems such as hormonal appetite regulators including ghrelin and leptin

• Stimulate digestion: The brain can send signals that stimulate digestive processes in the stomach, intestines and other parts of the digestive system, preparing the body to eat. In doing so, the brain initiates a response that involves activating the autonomic nervous system (responsible for automatic body functions, including the digestive system) and coordinating this response within the digestive system

• Metabolic regulation: The brain influences the body's metabolism by controlling how quickly food is processed and how energy is distributed.

• Induction of hunger or satiety: The brain can induce feelings of hunger or satiety depending on the body's current needs, food information, and other factors.

Based on the above, a model was built (in the form of a semantic network) of how a human feels hunger (Fig. 2).
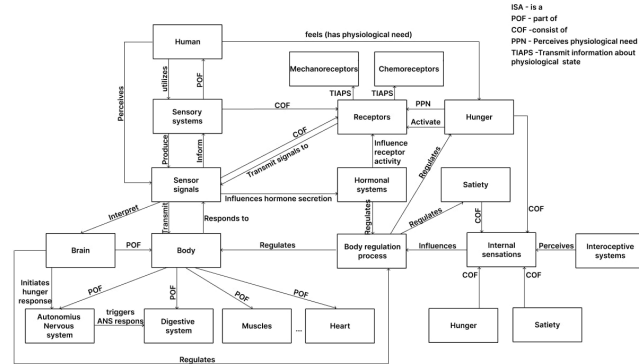

Figure 2. How a human feels hunger.

Discussion and Conclusion

What will we do with such lists? We can use the list to look for senses to include in cognitive architectures. First, there are many more senses than five! There are some external senses that can extend our cognitive and physiological (Hester et al., 2011) architectures. The internal senses are basically not included in cognitive architectures, but some are in agent and robots.

Kotseruba and Tsotsos (2020; in press) provide an extensive review of 84 cognitive architectures and defined the nomenclature of sensory modalities the architectures use. Here is their list, ordered in descending order of the number of modalities used: vision, symbolic input, proprioception, other sensors, audition, touch, smell and multi-modal.

External senses have tended to be more difficult to model than cognition; they require transducers to the real world (or models of them) and may require more knowledge and processing than cognition does. Including further external senses will also require tasks that use them. Most tasks studied to far do not use touch, smelling, hearing, or taste, partly because they are complicated to model and because psychology tends not to study them as often as the distal ones.

The internal senses are not often used in classic psychology studies. For example, only when we create longer running models (e.g., driving a great distance, Wu, Bagherzadeh, Ritter, & Tehranchi, 2023), will hunger, thirst, and voiding-related senses be necessary. Time estimation is a task, but it is not often used. It is used in more complex tasks that have not yet been modelled but is probably ubiquitous in all behavior.

The list of senses also suggests types of cognitive knowledge that are missing, for example, adjusting your body can reduce pain, or that a smell can be followed in the same way that active vision can lead to further information (Findlay & Gilchrist, 2003). How stress is perceived could be part of this story as well or for sense interaction.

When we consider the significance of this list of sensory modalities for cognitive architecture, it becomes evident that our understanding of sensory perception is constantly evolving. Recent discoveries, such as the identification of a sixth taste modality linked to lipid perception (Besnard, et al., 2016), complement traditional notions of sensory processing. Research on animal sensory systems also aids in designing and validating models for humans. For instance, insights into the sexual dimorphism of the olfactory system in mammals (Samaulhaq et al., 2008) corroborate findings (Oliveira-Pinto et al., 2014) showing that human dimorphism is conditioned by feminine characteristics. This, in turn, may explain the superior performance of women compared to men in olfactory tests.

Additionally, a number of studies have examined the intricate relationship between senses such as emotions and smells, as well as balance and nausea. Smells, just like emotions, can elicit positive, negative, or neutral reactions and influence our perception and behavior. This suggests common neural substrates underlying these phenomena.

The five senses have been mapped to brain regions. This larger list suggests that there are further regions to be assigned to the further senses. This mapping will help explain why we have such a big brain and how we use it.

This review shows that it is a big world out there still, and a big world even within ourselves yet to be modelled. We will need multiple sensors and multiple tasks to exercise these sensors.

References

Armstrong, S. A., & Herr, M. J. (2024). Physiology, Nociception. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.

Bach, J. (2008). *Principles of synthetic intelligence: Building blocks for an architecture of motivated cognition.* New York, NY: OUP.

Besnard, P., Passilly-Degrace, P., Khan, N.A. (2016). Taste of fat: A sixth taste modality? *Psych. Rev. 96*(1):151-76.

Buijs, T. J., & McNaughton, P. A. (2020). The role of cold-sensitive ion channels in peripheral thermosensation. *Frontiers in Cellular Neuroscience, 14,* [262].

Byrne, M. D. (2001). ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies, 55*(1), 41-84.

Dörner, D., & Güss, C. D. (2013). PSI: A computational architecture of cognition, motivation, and emotion. *Review of General Psychology, 17*(3), 297-317.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active Vision: The psychology of looking and seeing*. Oxford, UK: OUP.

Hester, R. L., Brown, A. J., Husband, L., Illescu, R., Pruett, D., Summers, R., et al. (2011). HumMod: A modeling environment for the simulation of integrative human physiology. *Frontiers in Physiology, 2*, Article 12.

Kirschfeld, K. (1976), The resolution of lens and compound eyes, In Zettler, F., & Weiler, R. (eds.) *Neural Principles in Vision,* 354–370.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review, 53*(1), 17–94.

Kotseruba, I., & Tsotsos, J. K. (in press). *The computational evolution of cognitive architectures*. Oxford, UK: OUP.

Kumar, C. M., & Van Zundert, A. A. J. (2018). Intraoperative Valsalva maneuver: A narrative review. *Can J Anaesth. 65*(5): 578-585.

Leite, I., Martinho, C., Pereira, A., Paiva, A. (2009). As time goes by: Long-term evaluation of social presence in robotic companions, *IEEE Int. Symp.on Robot and Human Interactive Communication, RO-MAN.* 669-674.

Liang, M., Damien, Ch., Fouad, B., Wei Zh. (2009). A new simple dynamic muscle fatigue model and its validation. I*nternational J. of Industrial Ergonomics*, *39*(1), 211-220.

Lindsay, P. H., & Norman, D. A. (1972). *Human information processing: An introduction to psychology.* Academic.

Lisina, M. I. (1997). *Общение, личность и психика ребенка* [Communication, personality and psyche of the child]. Moscow; Voronezh.

Mathews, Z., Lechon, M., Calvo, J. M. B., Duff, A. D. A., Badia, S. B. I., & Verschure, P. F. M. J. (2009) Insect-like mapless navigation based on head direction cells and contextual learning using chemo-visual sensors. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2243–2250

Mayer, E. A., Nance, K., & Chen, S. (2022). The gut–brain axis. *Annual Review of Medicine, 73,* 439-453.

Oliveira-Pinto A. V., Santos, R. M., Coutinho, R. A., Oliveira, L. M., Santos, G. B., Alho, A. T., et al. (2014). Sexual dimorphism in the human olfactory bulb: Females have more neurons and glial cells than males. *PLoS One. 5*;9(11): e111733.

Pelizza, L., & Ferrari, A. (2009). Anhedonia in schizophrenia and major depression: State or trait? *Ann Gen Psychiatry. 8*:22.

Ritter, F. E., Brener, M., & Bolkhovsky, J. B. (2021). An initial description of capabilities and constraints for a computational auditory system (an artificial ear) for cognitive architectures. *Proceedings of the Ninth Annual*

Patzelt, H., & Shepherd, D. A. (in press). A fatigue model of social venturing. *Small Bus Econ.*

Conference on Advances in Cognitive Systems, ACS-21_paper_11.

Rydström, A. & Bengtsson P. (2007). Haptic, visual and cross-modal perception of interface information. In D. de Waard, G. R. J. Hockey, P. Nickel, & K. A. Brookhuis (Eds.), *Human factors issues in complex system performance.* 399-409. Maastricht: Shaker Publishing

Samaulhaq, M., Tahir, Kh., Lone, Kh. (2008) Age and gender related differences in olfactory bulb glomeruli in human. *Biomedica. 24*. 12-27.

Sciutti, A., Mara, M., Tagliasco, V., Sandini, G. (2018). Humanizing human-robot interaction: On the importance of mutual understanding, *IEEE Technology and Society Magazine, 37*(1). 22-29.

Sekuler, R., & Blake, R. (2001). *Perception*. New York, NY: McGraw-Hill.

Shapiro, S. C., Kandefer, M. (2005). A SNePS approach to the Wumpus World agent or Cassie meets the Wumpus. *IJCAI-05 Workshop on Nonmonotonic Reasoning, Action, and Change (NRAC'05):* working notes.

Sneddon, L. U. (2018). Comparative physiology of nociception and pain. *Physiology 33*(1): 63-73.

Soudry, Y., Lemogne, C., Malinvaud, D., Consoli, S.-M., & Bonfils, P. (2011). Olfactory system and emotion: Common substrates, *Eur Ann Otorhinolaryngol Head Neck Dis, 128*(1), 18-23.

Stine, M. M., Klein, L. C., & Yatko, B. R. (2001). Daily caffeine use alters time perception. *Annals of Behavioral Medicine, 23*, S148.

Stoyanov, G. Moneva, K., Sapundzhiev, N et al. (2016). The vomeronasal organ—Incidence in a Bulgarian population. *J Laryngol Otol 130*: 344–347.

Taatgen, N., Van Rijn, H., Anderson, J. R. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psych. Rev. 114*(3), 577-598.

van Hartevelt, T. J., & Kringelbach, M. L. (2015). The olfactory cortex. In A. W. Toga (ed.), *Brain mapping,* Academic Press, 347-355.

Wang, C. X., Hilburn, I. A., Wu, D.-A., Mizuhara, Y., Cousté, C. P., Abrahams, J. N. H., Bernstein, S. E., Matani, A., Shimojo, S., & Kirschvink, J. L. (2019). Transduction of the geomagnetic field as evidenced from alpha-band activity in the human brain. *eNeuro 6*:0483–0418.

Wittmann, M. (2009). The inner experience of time. *Philos Trans R Soc Lond B Biol Sci. 364*(1525):1955-67.

Wu, S., Bagherzadeh, A., Ritter, F. E., & Tehranchi, F. (2023). Long road ahead: Lessons learned from the (soon to be) longest running cognitive model. *Proceedings of 21st International Conference on Cognitive Modeling* (ICCM), 281-287.

Zsolt, R. (2018). Chapter 4 - Fundamentals of strength training, Z. Radák, (ed.) *The physiology of physical training,* 55-80. Academic Press.

A Proposal for Extending the Common Model of Cognition to Emotion

Paul S. Rosenbloom[1] (Rosenbloom@USC.edu), John E. Laird[2] (John.Laird@cic.iqmri.org), Christian Lebiere[3] (CL@CMU.edu), Andrea Stocco[4] (Stocco@UW.edu), Richard H. Granger[5] (Richard.Granger@Dartmouth.edu) & Christian Huyck[6] (C.Huyck@MDX.ac.uk)

[1] Institute for Creative Technologies & Thomas Lord Dept. of Computer Science, University of Southern California

[2] Center for Integrated Cognition, IQMRI

[3] Department of Psychology, Carnegie Mellon University

[4] Department of Psychology, University of Washington

[5] Department of Psychological and Brain Sciences, Dartmouth University

[6] Department of Computer Science, Middlesex University

Abstract

Cognition and emotion must be partnered in any complete model of a humanlike mind. This article proposes an extension to the Common Model of Cognition – a developing consensus concerning what is required in such a mind – for emotion that includes a linked pair of modules for emotion and metacognitive assessment, plus pervasive connections between these two new modules and the Common Model's existing modules and links.

Keywords: Common Model of Cognition; emotion; metacognitive assessment; cognitive architecture

Introduction

The *Common Model of Cognition* (Rosenbloom, Lebiere & Laird, 2022) – née the *Standard Model of the Mind* (Laird, Lebiere & Rosenbloom, 2017) – is a developing consensus concerning what must be in a cognitive architecture to support *humanlike minds*. The consensus is derived from existing cognitive architectures, from researchers who study them, and from results relevant to them, with humanlike minds comprising human minds plus any other natural or artificial minds similar enough to be modellable in the same manner at the chosen level of abstraction.

The Common Model is not intended to be a cognitive architecture in the traditional sense, in being abstract, radically incomplete, and not directly executable. What it includes is limited to what the community can reach a consensus on concerning its necessity for humanlike cognition. Sufficiency considerations play into what topics are considered for consensus building but play no direct role in judging what is actually to be included.

This article reports on an effort to address one major source of incompleteness in the Common Model – concerning *emotion* – that has not yet reached a consensus with respect to necessity. It thus amounts to a proposal for how to extend the Common Model to particular aspects of emotion but not (yet) an actual extension of the Common Model to emotion.

As stated in Larue et al. (2018), "Modeling emotion is essential to the Common Model of Cognition … because emotion can't be divorced from cognition. … Emotions play an important functional role, with the purpose of helping us to survive and adapt in complex and potentially hazardous physical and social domains (Panksepp & Biven, 2012). They

aren't necessarily finely tuned but guide our behavior in directions evolution has taught us are wise."

What is proposed here is nowhere near a full model of emotion. It focuses on only the architectural aspects of how emotional states arise and affect cognition; and this it only does abstractly, not delving into the details of appraisal and dimensional models. It also has nothing to say at this point about such topics as how emotional states are reflected in external expressions. Still, the intent is to take a significant step in considering how emotion relates to architectures that align with the Common Model.

The next section provides background on the Common Model and how we arrived at this proposal. The subsequent two sections provide more details on two new modules that are proposed for inclusion into the Common Model – one for *emotion* and one for *metacognitive assessment* – and how they interact with the rest of the model. The final section summarizes what has been proposed here.

Background
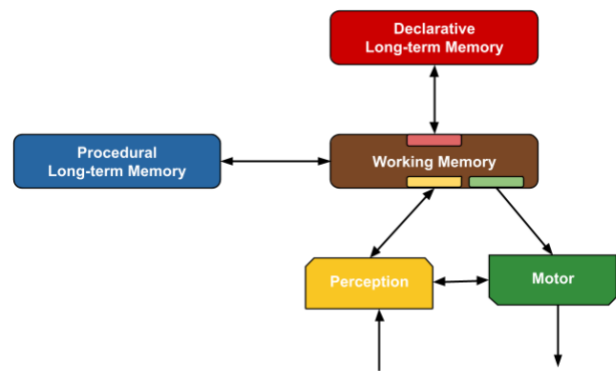
Figure 1: The Common Model of Cognition.

Figure 1 shows the basic structure of the Common Model. It comprises a central working memory, two long-term memories, and perception and motor modules. Working memory (WM) represents the current situation. Procedural long-term memory, which consists generically of rule-like structures, has direct access to all of WM. The other modules interact with it through dedicated buffers. Declarative long-term memory here does not (yet) distinguish between

semantic and episodic knowledge. The perception and motor modules are minimally defined.

This figure is accompanied by sixteen assumptions about how it all works, divided up according to whether they bear on: (A) structure and processing; (B) memory and content; (C) learning; or (D) perception and motor control. Key assumptions, for example, include: (A3) there is significant parallelism both within and across the modules; (A4) sequential behavior arises from a cognitive cycle operating at ~50 msec in humans; (B1) long-term memories contain symbolic data with associated quantitative metadata; (B2) global communication occurs via WM; and (C2) learning occurs incrementally as a side effect of performance.

A broad survey of cognitive architectures can be found in Kotseruba and Tsotsos (2020), including examples of architectures with aspects of emotion. The proposal here, however, grew more directly out of an earlier analysis of the relationship of emotion to the Common Model (Larue et al., 2018) that later fed into a virtual workshop on the topic in June 2022. Out of the final session of that workshop came an initial consensus (Figure 2).

Figure 2: A version of the initial CMC+Emotion synthesis from the 2022 workshop.

The core of this figure is the pre-existing Common Model. Added to it is an *affect* module intended to capture key aspects of emotion. It receives input from physiology, to react to bodily state; from perception, to invoke immediate reactions to the state of the world without requiring direct cognitive participation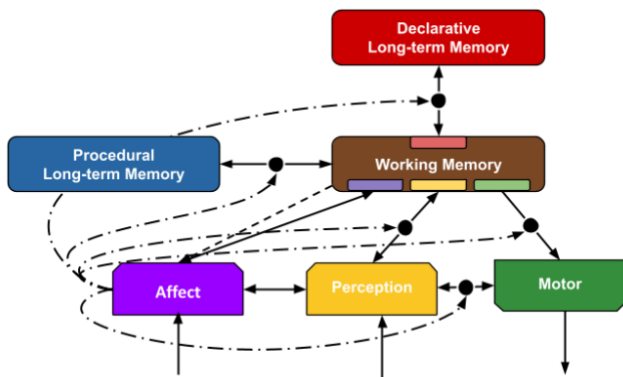; from its WM buffer, to invoke reactions to the state of the cognitive system; and from WM more broadly, and likely more diffusely, to set the context for affective processing. In return, the affect module provides input to its WM buffer, to enable reasoning by the cognitive system about its results; it filters communications between the other modules and WM; and it directly affects perception (see, e.g., Zadra & Clore, 2011).

In this model, reasoning about emotion – that is, the "cold" aspects of emotion, whether concerning oneself or others – is presumed to occur via standard cognitive processing within the original modules of the Common Model. The higher-level aspects of appraisal theory (e.g., Marsella & Gratch, 2009; Moors et al., 2013; Scherer, 2001) would thus fit here. The "hot" aspects of emotion are presumed to be the province of

the physiological system that is not shown but that provides input to the affect module (e.g., Dancy, 2013). The model provides scope for both of these aspects of emotion but has nothing further to say about either of them.

What was termed the "warm" aspects of emotion at the workshop are the architectural facets of emotion processing, such as processing within the affect module itself, how metacognitive assessment lays the groundwork for it, and its impact on the rest of the Common Model. As with cold and hot emotion, the model does not delve into the internals of these two warm modules, but it does propose extending the Common Model's architecture to include them. Still, given that warm emotion is architectural, it would make sense for future work on the Common Model to accrete further details about how they operate.

One example architectural precedent for key aspects of this proposal can be found in the Sigma cognitive architecture (Rosenbloom, Gratch & Ustun, 2015). Attention there differentially abstracts messages throughout the cognitive system, affecting both communication across modules and within them, driven by a combination of two low-level, architecturally computed appraisals – *desirability* and *surprise* – which are themselves a function of what is perceived, what is learned, and what is in WM. The results of these appraisals then arrive back in WM. Other examples include West and Young's (2017) proposal for a similar extension to the Common Model that accesses WM like procedural long-term memory while providing subsymbolic evaluations back to both long-term memories, and Smith et al.'s (2021) argument for activation in ACT-R to include emotion via an additive scalar term.

Although there was a sense of consensus coming out of that workshop with respect to something like Figure 2, it was the result of only a few days' work by a subset of the community that did not actually come together until the final session. It thus did not seem right to consider it by itself as an official consensus. And, even if it were to become such, and thus part of the Common Model, it was quite minimal. So, the first four co-authors on this article set out to push the model further before going back to the community to see if a more thorough consensus was reachable. The proposal here is the product of these deliberations.

Outline of the Proposal

Figure 3 outlines this new proposal. The changes in module locations from Figure 2 are purely cosmetic, to simplify and deconflict the resulting diagram. Each of the remaining changes reflects refinements that have been made to the initial CMC+Emotion model in Figure 2. This is a complex figure that will be broken down further in the next two sections.

One change that may look cosmetic is relabeling the affect module as *emotion*. Significant consideration went into the question of exactly what function this module – which has at various times been labeled affect, emotion, or physiology – should serve, as well as how it connects to the other modules. It being labeled emotion in Figure 3 implies this concluded

with the idea that while it senses physiology this module's role is to generate emotional vectors, such as <valence, arousal> or more extended vectors, that are central to dimensional models of emotion (e.g., Juvina, Larue & Hough, 2018; Mehrabian & Russell, 1974; Rubin & Talarico, 2009), but without a commitment to the size or contents of such a vector.
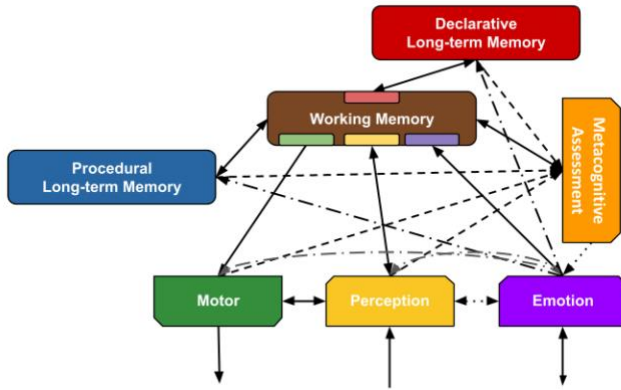

Figure 3: Refined CMC+Emotion synthesis that is the basis of the current proposal.

One major change was due to noticing that an important pathway, and associated module, was missing from Figure 2 that generates low-level appraisals based on observing the existing modules and the traffic among them, and transmits the results to the emotion module. This has taken the form here of a metacognitive assessment module that is discussed after the emotion module.

A second major change is that the emotion module here not only filters communications between other modules and WM but also affects how the modules themselves operate. One simple example of the latter is how emotions may yield rewards for reinforcement learning in procedural long-term memory (e.g., Marinier & Laird, 2008), or somatic markers in declarative long-term memory (Damasio, 1994).

More broadly, extensive evidence indicates that emotions can exert specific influences on memory storage and retrieval, affecting the processing of learned information. These include much-studied everyday effects such as state-dependent memory (e.g., Eich, 1995) and clinical effects such as post-event emotional and traumatic responses (Brewin, 2011) that are directly mediated via cortico-amygdala loops (e.g., Fadok et al., 2018; Grundemann et al., 2019). Several specific hypotheses about the nature of the effects of emotional content on memory storage, retrieval, and inference can be posed both behaviorally and neurally to clarify features of emotion-memory interactions (e.g., Fadok et al., 2018; Janak & Tye, 2015; Kesler, 2001; Saarimaki, 2016).

The emotion and metacognitive assessment modules can both be seen as analogous to the perception module, although they perceive physiology and the cognitive system respectively rather than the external environment. Similar analogies are also conceivable between these new modules and the motor module when their ability to act on their environments is considered.

The beginnings of an attempt has been made to determine if a consensus was reachable around this new proposal, involving two separate emails to the workshop attendees requesting input from them on it, one informal, as free text, and one structured more formally as a questionnaire, but this process proved insufficient to yield a consensus even among the workshop attendees, so no attempt has yet been made at achieving a broader community consensus.

This material is therefore presented as a proposal for further consideration rather than as an agreed-upon extension to the Common Model. The other two co-authors on this paper were workshop attendees who agreed to join in this latest stage of proposal refinement and presentation.

Emotion Module

Figure 4 is a simplified version of Figure 3 that eliminates metacognitive assessment. This version is much like Figure 2, but for two key extensions. First, there is a connection from the emotion module back down to physiology for emotion vectors to affect physiology; for example, when a cognitively identified threat – such as a verbal threat – requires the body to prepare to respond. Second, the dot-dash arrows from the emotion box now point to the junctions between the non-WM Common Model modules and their links to WM. This is to indicate that not only can the vectors from the emotion module filter communication along these links, but they can also modify how these modules work. Although the model does not specify how this happens, examples may include altering how procedural memory selects actions to execute and how declarative memory determines what knowledge to retrieve.

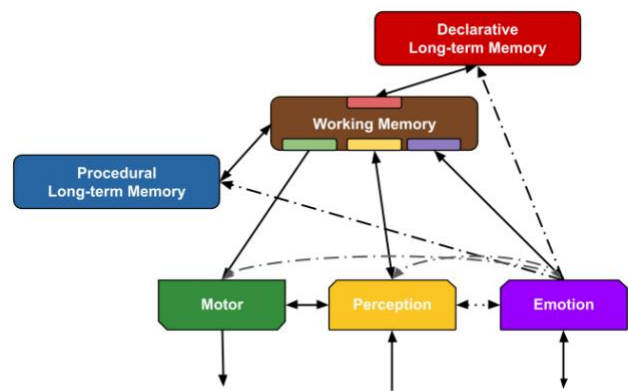Figure 4: A version of Figure 3 simplified to omit metacognitive assessment.

In Figure 4, as in Figure 2, there is a WM buffer associated with the emotion module. The arrow to this buffer enables the vectors generated by the module to appear in working memory. In Common Model fashion, these vectors could arrive as symbolic data identifying the dimensions of the vector with quantitative metadata that specifies their values.

The return arrow provides cognitive input to the emotion module, including high-level appraisals. As discussed further in the next section, low-level appraisals follow a different route, via the metacognitive assessment module. With input to the emotion module that includes both high-level and low-level appraisals, and output from it in the form of vectors, this proposal implies a connection between appraisal and vector models in which the former are inputs to creation of the latter. However, vector generation also depends on inputs from physiology and perception, and possibly other content in WM.

Two simple examples of how this might proceed based just on appraisals, albeit at opposite ends of the vector-length spectrum, are (1) generation of large vectors by assigning one slot to each appraisal; or (2) generation of intensity and valence pairs by aggregating over the values of all appraisals for the former and differentially over positive and negative appraisals for the latter. Under the second option, physiological inputs might combine straightforwardly with the two values derived from appraisals.

Figure 4 includes a bidirectional arrow between emotion and perception, although it is shown as a dotted line because it remains unclear whether a direct return path from emotion to perception is needed in addition to the curved arrow that already indicates emotional modulation of perception.

Metacognitive Assessment Module

In attempting to include appraisals and their relationship to emotions, we ended up modeling them generically as aspects of metacognition – where the cognitive system operates on itself – a large-scale topic of its own on which an overall consensus has not yet been reached with respect to the Common Model, although Kralik et al. (2018) did begin exploring this question.

The particular point of interest here is that some forms of low-level appraisals, such as surprise and familiarity, can be thought of as metacognitive assessment that is grounded in fixed, architectural sensors that observe what is happening within the overall cognitive system. As one simple example, both surprise and familiarity are computed architecturally in Sigma based on monitoring its learning process.

This approach would put such appraisals in the same category as, for example, a sensor for *feeling of knowing* that assesses when declarative memory will be able to retrieve an appropriate memory given a cue (Nhouyvanisvong & Reder, 1998). It would also put them in the class of "warm" aspects of emotion.

Figure 5 shows a version of Figure 3 that includes the metacognitive assessment module for low-level appraisals, but which is simplified via the removal of the dot-dash arrows from the emotion module to the relevant Common Model modules.

In this figure it can be seen how the low-level appraisals from this new module act as inputs to the emotion module. However, it remains unclear in general whether these appraisals should arrive directly from the metacognitive assessment module via the dotted arrow between the two

modules, or whether this path can be omitted given the existence of the path via solid arrows that traverses WM.
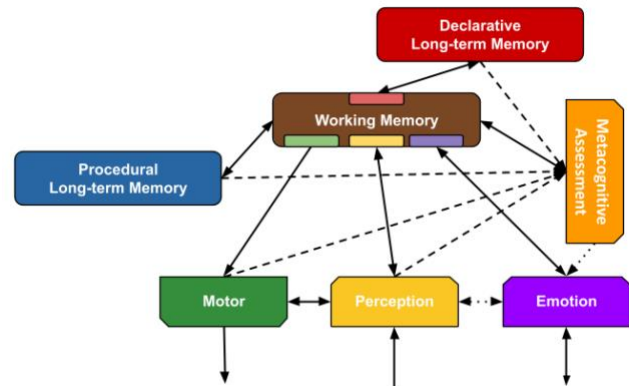
Figure 5: A version of Figure 3 simplified to omit the dot-dash arrows from the emotion module to the non-WM Common Model modules.

High-level "cold" appraisals, such as causal attribution, are considered to be essentially cognitive in nature, although Figure 5 abstracts over whether this form of metacognition occurs within (possibly a recursion on) the same cognitive system (e.g., Rosenbloom, Laird & Newell, 1988) or via a distinct metacognitive system (e.g., Cox, Oates. & Perlis, 2011; Sun, Zhang & Matthews, 2006). It makes sense to defer attempting to resolve such a question until a full exploration is begun of how to extend the Common Model to metacognition.

As shown in Figure 5, the metacognitive assessment module receives input from each of the other non-WM modules in the original Common Model. That these connections are from junctures between these other modules and their links to WM is intended to indicate that the metacognitive assessment module can sense both their communication with WM and what is going on within them, although the figure abstracts over exactly what is sensed. None of this sensing of the cognitive system is reflected in Figure 2, but it is intended to effectively be the inverse of how the emotion module acts upon these junctures.

There are no arrows back from the metacognitive assessment module to the other modules, which might be expected in a full analysis of metacognition, but these do not appear necessary for emotional metacognitive assessment. Instead, what feedback does occur goes through the emotion module before it reaches them. Whether direct backward connections are ultimately needed in addition to this route through the emotion module remains to be seen.

Although metacognitive assessment is shown in Figure 5 as a separate module, it is not yet clear whether it should truly be considered a module on its own versus there being merely bits of it distributed across the other modules and connections, where it is presumed that the sensing actually occurs. It is shown as a module here to leave open the possibility of architectural across-module appraisals – such

as the earlier Sigma example in which attention is based on both surprise and desirability – rather than assuming that all low-level assessments are specific to one module or that all combinations of them happen cognitively. However, the necessity of such a possibility remains as another open question.

An arrow is shown in Figure 5 from metacognitive assessment to WM to make low-level appraisals accessible to cognition in support of higher-level appraisal, including complex across-module appraisals, as well as other relevant cognitive processing. The reverse arrow, from WM back to the module indicates WM affecting metacognitive assessment.

It is left open whether these interactions are as unconstrained as those between procedural long-term memory and WM or whether they are constrained to go through a module-specific buffer, as is the case with the other modules. However, if the arrow from WM to the metacognitive assessment module is unconstrained, it may be able to substitute for Figure 2's arrow from WM to the affect module, supporting a flow from all of WM, through metacognitive appraisal, to emotion.

The previous section raised the question of whether a direct connection is required from emotion to perception. In the current context, the existence of a pathway from perception, through metacognition, to emotion raises the reverse question, as to whether a direct link from perception to emotion is necessary when this slightly less direct path already exists.

Summary

The proposal presented here for extending the Common Model of Cognition to aspects of emotion includes a new emotion module that can affect the workings of the existing non-WM modules as well as filter their communications with WM. It also includes a new metacognitive assessment module that can perceive WM plus the workings of the existing non-WM modules and their communications with WM. These two modules, plus links between them and between the emotion module and physiology, comprise the core of the proposed model, as outlined in Figure 3.

This model is of course incomplete in many ways with respect to the full complexity of emotion. It shares the Common Model's natural abstraction and incompleteness in terms of only including aspects about which there is a consensus, although here this is in terms of what might become a consensus. Thus, this article still reflects only a beginning of a beginning at extending the Common Model to emotion, even while building on multiple earlier efforts in this direction. Still, given the importance of the connection between cognition and emotion, it hopefully provides a basis for a wider discussion of what should be added to the Common Model in support of extending it to emotion.

References

Brewin C. (2011). The nature and significance of memory disturbance in posttraumatic stress disorder. *Annual Review of Clinical Psychology*, *7*, 203-227.

Cox, M.T., Oates, T., & Perlis, D. (2011). Toward an integrated metacognitive architecture. *AAAI fall symposium: Advances in cognitive systems*.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Penguin Books.

Dancy, C. L. (2013). ACT-RΦ: A cognitive architecture with physiology and affect. *Biologically Inspired Cognitive Architectures*, *6*, 40-45.

Eich E. (1995). Searching for mood dependent memory. *Psychological Science*, *6*, 67-75.

Fadok, J., Markovic. M/, Tovote. P/, & Lüthi, A. (2018). New perspectives on central amygdala function. *Current Opinion in Neurobiology*, *49*, 141-147.

Gründemann, J., Bitterman, Y., Lu, T., Krabbe, S., Grewe, B., Schnitzer. M., & Lüthi. A. (2019). Amygdala ensembles encode behavioral states. *Science*, *364*, eaav8736.

Janak. P., & Tye, K. (2015). From circuits to behaviour in the amygdala. *Nature*, *517*, 284-292.

Juvina, I., Larue, O., & Hough, A. (2018). Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. *Cognitive Systems Research, 48,* 4–24.

Kesler, M., Andersen, A., Smith, C., Avison, M., Davis, C., Kryscio, R., & Blonder, L. (2001). Neural substrates of facial emotion processing using fMRI. *Cognitive Brain Research*, *11*, 213-226.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*, 17-94.

Kralik, J. D., Lee, J. H., Rosenbloom, P. S., Jackson, P. C., Epstein, S. L., Romero, O. J., Sanz, R., Larue, O., Schmidtke, H. R., Lee, S. W., & McGreggor, K. (2018). Metacognition for a Common Model of Cognition. *Procedia Computer Science*, *145*, 740-746.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A Standard Model of the Mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*, 13-26.

Larue, O., West, R., Rosenbloom, P. S., Dancy, C. L., Samsonovich, A. V., Petters, D., & Juvina, I. (2018). Emotion in the Common Model of Cognition. *Procedia Computer Science*, *145*, 730-739.

Marinier, R., & Laird, J. (2008). Emotion-driven reinforcement learning. In *Proceedings of the 30th annual conference of the cognitive science society*.

Marsella, S., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research,10*, 70-90.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*, 119-124.

Nhouyvanisvong, A., & Reder, L. M. (1998). Rapid feeling-of-knowing: A strategy selection mechanism. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: Cognitive and social dimensions*. Sage Publications, Inc.

Panksepp, J., & Biven, L. (2012). *The archaeology of mind: Neuroevolutionary origins of human emotions*. New York, NY: W. W. Norton & Co.

Rosenbloom, P. S., Gratch, J., & Ustun, V. (2015). Towards emotion in Sigma: From appraisal to attention. In *Proceedings of the 8th conference on artificial general intelligence* (pp. 142-151).

Rosenbloom, P. S., Laird, J. E., & Newell, A. (1988). Meta-levels in Soar. In P. Maes, & D. Nardi (Eds.), *Meta-level architectures and reflection* (pp. 227-240). Amsterdam, Netherlands: North Holland.

Rosenbloom, P. S., Lebiere, C., & Laird, J. E. (2022). Cross-pollination among neuroscience, psychology and AI research yields a foundational understanding of thinking. *The Conversation*.

Rubin, D. C. & Talerico, J. M. (2009). A comparison of dimensional models of emotion. *Memory*, *17*, 802-808.

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I., Lampinen, J., Vuilleumier, P., Hari, R., Sams, M., & Nummenmaa, L. (2016). Discrete neural signatures of basic emotions. *Cerebral Cortex*, *26*, 2563-2573.

Scherer, K. (2001). Appraisal considered as a process of multilevel sequential checking. In K. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research*. New York: Oxford University Press.

Smith, B. M., Thomasson, M., Yang, Y. C., Sibert, C., & Stocco, A. (2021). When fear shrinks the brain: A computational model of the effects of posttraumatic stress on hippocampal volume. *Topics in Cognitive Science*, *13*, 499-514.

Sun, R., Zhang, X., & Mathews, R. (2006). Modeling meta-cognition in a cognitive architecture. *Cognitive Systems Research*, *7*, 327-338.

West, R. L., & Young, J. T. (2017). Proposal to add emotion to the Standard Model. *Common Model of Cognition Bulletin*, *1*, 487-492.

Zadra J. R., & Clore G. L. (2011). Emotion and perception: The role of affective information. *WIREs Cognitive Science*, *2*, 676-685.

How Working Memory Influences Knowledge Reconstruction in Collaborative Learning: Investigation Using Human Experimentation and ACT-R Simulation

Shigen Shimojo (sshimojo@fc.ritsumei.ac.jp)
Ritsumeikan Global Innovation Research Organization, 2-150 Iwakura-cho
Ibaraki-shi, Osaka, Japan

Yugo Hayashi (yhayashi@fc.ritsumei.ac.jp)
College of Psychology, Ritsumeikan University, 2-150 Iwakura-cho
Ibaraki-shi, Osaka, Japan

Abstract

Collaborative learning is when learners reconstruct one's knowledge based on others' knowledge and then gain understanding. However, it is indicated that working memory inhibits the cognitive process learners acquire and use other's knowledge. Additionally, it is challenging to manipulate learners' working memory and capture the cognitive process during collaborative learning in psychological experiments. Therefore, this study investigated how working memory influenced the search for knowledge reconstruction in a psychological experiment and further examined the nature of the cognitive process using ACT-R, a cognitive architecture and theory for human cognition. Both laboratory experiments and simulations revealed a positive correlation between working memory and correct. Also, those revealed no correlation between working memory and incorrect. In model-based simulations, we found that reconsidering based on others' knowledge was effective when working memory was high. This study contributed to developing pedagogical agents as collaboration members and teachable agents to support collaborative learning.

Keywords: Collaborative learning; Working memory; Computer simulation; ACT-R

Introduction

The efficacy of collaboration is that learners reconstruct their ideas based on other people's knowledge and ideas (Shirouzu & Miyake, 2002). What learners externalize based on others' ideas leads to correct knowledge use (Chi & Wylie, 2014). However, learners contribute less to collaborative learning than to individual knowledge; this is called "collaborative inhibition." In collaboration, people find it challenging to acquire others' perspectives (Hayashi, 2020). According to Shimojo and Hayashi (2019), visualizing each other's knowledge using concept maps facilitates the use of correct knowledge. Correct knowledge is used so learners can use the correct knowledge in the text. To fill in the learning text in a link to concept maps, learners must understand the learning text. Therefore, acquiring others' knowledge facilitates the use of correct knowledge.

Considering these facts, this study investigates how learners use correct knowledge in a collaborative learning environment using concept maps based on the human cognitive process. Additionally, this study developed a cognitive model to explain how learners retrieve their knowledge and use correct and incorrect knowledge in explanation activities to investigate these mechanisms. Subsequently, this paper explains the factors involved in using knowledge to learn, overview theory, and cognitive model of collaborative learning.

Factor on Knowledge Use During Collaboration

It is commonly believed that learning and individual memory are closely related, particularly to working memory. There is a dual-store model as a theory of memory, which indicates elaboration by rehearsal (Atkinson & Shiffrin, 1968). The cognitive process by which information is stored in long-term memory is revealed. However, it has been shown that collaboration prevents memories from being stored or recalled. For example, collaborative memory retrieval results in fewer than nominal pairs that learners don't interact (Sjolund, Erdman, & Kelly, 2014). It is most likely that collaboration not only improves memory but also forgetting and error of retrieval (Rajaram & Pereira-Pasarin, 2010). Moreover, cognitive load inhibits collaborative learning, called Collaborative Cognitive Load Theory (CCLT; (Janssen & Kirschner, 2020). The cognitive load theory is also based on working and long-term memory (Schweppe & Rummer, 2014). The working-memory store feature provides information that learners learn new and activated long-term memory temporarily (Cowan, Morey, & Naveh-Benjamin, 2020). Therefore, it plays an essential role in working memory when investigating the effects and inhibitory factors of collaborative learning.

In a collaborative learning environment in which students reconsider their knowledge based on others, the cognitive load is higher, and then it is necessary to consider how working memory is involved. Indeed Engelmann and Hesse (2010) indicates that acquiring others' knowledge is inhibited by working memory. However, as shown in the Interactive-Constructive-Active-Passive (ICAP) theory, correct knowledge is facilitated by engaging in the deep cognitive process of reconstructing one's knowledge based on others' knowledge (Chi & Wylie, 2014). Consequently, working memory and reconstruction of one's knowledge based on others' knowledge are likely related to the effect of collaborative learning. Therefore, this study investigates the influence of working memory on using the correct knowledge and reconstructing one's knowledge based on others, as well as the reason reconstruction of one's knowledge based on others' knowledge is effective. This study adopted model-based approaches to address these research questions.

Cognitive Models of Collaborative Learning

Collaborative learning is effective in terms of memory, but it has not yet been shown why learners outperform through in-

teraction in collaboration based on cognitive process. Most previous studies have investigated theories and models that are not based on the cognition of humans or theories based on cognition rather than cognitive models based on human thinking and memory. For example, transactive memory refers to a set of individual memory systems combined with the communication between each individual group member (Wegner, 1987). Moreover, collaborative learning involves divergence and convergence focused on the essential memory function of individual working memory (Jorczak, 2011). An individual's working memory is involved in collaborative learning because when externalizing the representation of knowledge stored in memory, the working memory needs to process related information from long-term memory. The cognitive process in collaborative learning is that the externalized information is divergent and convergent. Group members store divergent information, and learners identify and select relevant information, which is a convergence. However, misconceptions are generated not only by relevant information during divergence information processing (Jorczak, 2011).

In cognitive science, cognitive models have been developed and simulated, but they don't focus on collaborative learning. Adaptive Control of Thought-Rational (ACT-R) architecture is designed and used as an intelligent tutoring system (Anderson, Corbett, Koedinger, & Pelletier, 1995). ACT-R is a cognitive architecture and theory for human cognition. Several studies have developed computational models for collaboration. Walker, Rummel, and Koedinger (2014) examined an adaptive collaborative learning system using a collaborative learning model by model trace, learner's knowledge trace algorithm, machine classification, and Cognitive Tutor Algebra classification. Hayashi and Koedinger (2019) investigated the cognitive process by which people share their differences and knowledge using cognitive task analysis. Suebnukarn and Haddawy (2006) explored individual and collaborative student clinical reasoning modeling to develop an intelligent tutoring system. This is an inferred model for medical treatment and a minimal reasoning model. However, these studies don't focus on human cognition and could not investigate working memory and reconstructing one's knowledge based on others' knowledge based on the cognitive process of humans. Therefore, a collaborative learning model needs the primary memory function of humans using ACT-R.

Consequently, this study investigates the influence of working memory on correct and incorrect knowledge and the relationship between reconstruction based on others' knowledge and working memory. First, we investigated the impact of working memory in a laboratory experiment and reproducing human results using simulation. Second, we investigated the use of correct and incorrect knowledge by observing the model's performance of the relationship between the reconstruction based on others' knowledge and working memory. At that time, ACT-R—a cognitive architecture—was used because it was easier to model working memory and the degree of memory retrieval based on the knowledge of others by ma-

nipulating parameters, and it has a visual buffer that reproduces human thinking and behavior in real situations.

Goal and Hypothesis

This study investigated the influence of working memory and the reconstruction of one's knowledge based on others' knowledge while creating a concept map in collaboration using laboratory experiments and simulation based on a model-based approach. The focus on individual working memory is to understand collaborative learning from the information processing perspective. Learning and memory are related, and learning is influenced by working memory, cognitive load, and knowledge activation. Therefore, if working memory is greater, learners may use correct knowledge because of the influence of working memory on the activation of knowledge. In addition, if working memory is greater, learners may use incorrect knowledge because of the activation of incorrect knowledge. This study hypothesized that working memory and the use of correct knowledge were related (H1). Additionally, working memory and the use of incorrect knowledge were associated (H2). In the simulation, we investigate the influence of working memory by manipulating the parameter on working memory in knowledge activation.

Method

Participants

The participants were 20 university students majoring in psychology (14 women and six men), with a mean age of 19.00 years (SD = 0.89). We also obtained informed consent about data confidentiality, anonymity, and withdrawal at any time in writing. Participants did not experience the experimental task of this study and knew the causal attribution they learned. This study was approved by the university's ethics committee, to which the author belongs.

Experimental Materials

We used learning text that learners read to learn the causal attribution of success and failure, episode, and created concept maps. Tool-created concept maps were designed collaboratively and synchronously. The episode in which a student talked about being anxious about a new semester was used in Weinberger and Fischer (2006). The learning text included information about internal-external, stable-unstable, and controllable-uncontrollable. Internal-external is the actor attributes a cause of a phenomenon to own or external of own. Stable-unstable is the actor that attributes the cause of the phenomenon to be stable or unstable over time. Controllable-uncontrollable is the actor attributes a cause or phenomenon to controllable or uncontrollable things and others.

Procedure

Learners first learned to create concept maps and conducted the 2-back task based on Dobbs and Rule (1989) second. In particular, learners filled in numbers before the n-2 trial in the n-trial using a keyboard. Third, learners referred to the learning texts and learned that. Fourth, learners referred to an

episode in which a student talked about being anxious about a new semester and then created concept maps individually by inferring why a person was anxious about the new semester (10 minutes). Fifth, learners referred to each other's concept maps and created them collaboratively (15 minutes). In particular, learners filled in nodes about anxiety and its cause and link causal attribution (e.g., anxious-internal-effort) . The sequence of the study was repeated. Figure 1 shows a screenshot of the collaboration. Concept maps include nodes and links: a node is about information about the episode, and a link is about learning the text.
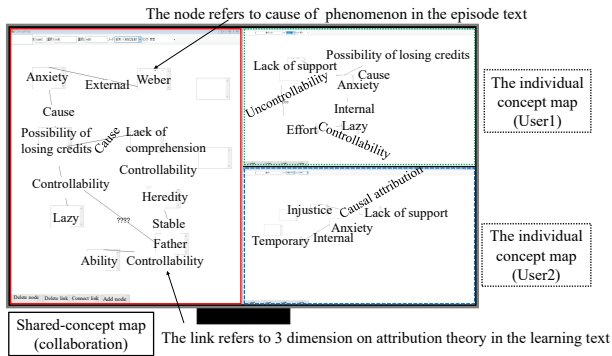

Figure 1: Capture of the screen during collaborative learning. The right-hand side shows two windows of concept maps generated in the individual phase. The left-hand side shows the shared window for creating collaborative concept maps

Measurement

This section explains the data used in this analysis. Working memory was adopted as the correct answer rate for the 2-back task. Correct knowledge was used in the number of links about the three dimensions in collaborative concept maps (correct links). Learners filled in the correct link about the three dimensions needed to understand the causal attribution of success and failure correctly. Incorrect knowledge was used for the number of incorrect links in three dimensions that were out of three dimensions and mistake links between nodes in collaborative concept maps (incorrect links). The correct link was knowledge of the correct or incorrect answer that did not exist on the individual concept map. The critical point is that the correct links are an indicator of the collaborative investigation result because of the effect of the collaborative discussion. Moreover, this indicator was adopted to investigate the influence of reconstructing one's knowledge based on others' knowledge of using the correct knowledge.

Result

First, we conducted an analysis using Pearson's correlation between the correct answer rate of the 2-back task and the number of correct links. Figure 2 shows a scatter plot of the correct answer rate and the number of links. A moderately positive correlation was observed ($r = .46, p < .05$). Consequently, H1 was supported, revealing that correct knowl-

edge was related to working memory. Next, we conducted a correlation analysis between the correct answer rate of the 2-back task and the number of incorrect links to test H2. Figure 3 shows a scatter plot of the correct answer rate and the number of incorrect links. No correlation was observed ($r = .06, p = .80$). H2 was not supported, revealing that incorrect knowledge was unrelated to working memory because selecting knowledge according to the task.
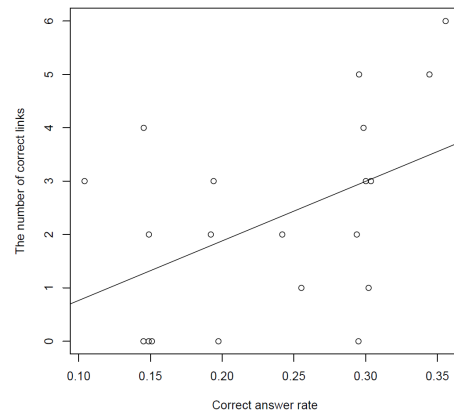
Figure 2: Relationship between the correct answer rate of the 2-back task and correct links
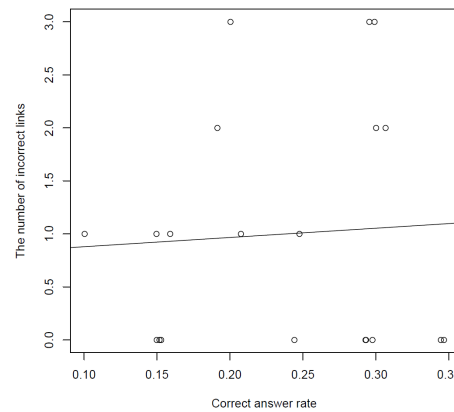
Figure 3: Relationship between the correct answer rate of the 2-back task and incorrect links

Simulation by ACT-R

Psychological experiments have shown a correlation between working memory and the number of correct links and not a correlation between working memory and the number of incorrect links. In other words, learners converged knowledge, selecting knowledge according to the task. In this section, a model using ACT-R was created and simulated. We clarify the relationship between working memory and reconsidering one's knowledge based on others' knowledge. It is challenging to reconsider one's knowledge based on the knowledge of others in laboratory experiments.

Using the ACT-R architecture (Anderson et al., 1995), we developed a model of knowledge use and examined the process of knowledge use during learning based on different types of knowledge based on Hayashi and Shimojo (2022a, 2022b). Hayashi and Shimojo (2022a) indicated perspective taking. Also, Hayashi and Shimojo (2022b) revealed retrieved knowledge based on others. The model replicates the task of creating a concept map in a collaborative task (15 min) and assumes that the knowledge of individual concept maps and the knowledge of the learning text exist as declarative knowledge. Figure 4 illustrates the flow of the model. The specific flow of the model is as follows: The learner acquires the character strings in the concept map("find-unattended-node1-indivi") and pays attention to the ones that still need attention("attend-node1-indivi") and add to declarative knowledge("encode-node1-indivi"). Learners repeat the same process for concept maps of one proposition (phenomenon, three dimensions, and causes), self, and others (from "find-unattended-node1-indivi" to "encode-node2-other"). They inscribe the declarative knowledge about one proposition, determine whether the partner's and knowledge are different or the same("individual-retrieve"), and if different, reconsider based on knowledge ("knowledge-retrieve-individual") or partner's knowledge("knowledge-retrieve-other") based on the utility value of procedural Knowledge. If the knowledge is identical (if correct), fill in the text(from "fill-in-node1" to "fill-in-node2"), or do not fill in the text("fail"). In "individual-retrieve," "knowledge-retrieve-individual," and "knowledge-retrieve-other," the model used knowledge activation.

Parameters

This simulation examined the relationship between working memory and correct knowledge use. (1) expresses knowledge activation. (2) expresses the noise value. B_i is the chunk base level, W_{kj} is the weighting, and S_{ji} is a measure of association strength. ε is the noise. In Matsumuro et al. (2018), W in the following activity value of (1) is employed as the operating parameter of the working memory. Therefore, we manipulated the W in (1 to simulate the effect of working memory. The parameters were set to 0.7, 1, and 1.3.

$$A_i = B_i + \Sigma_k \Sigma_j W_{kj} S_{ji} + \varepsilon \qquad (1)$$

$$\sigma^2 = \frac{\pi^2}{3} S^2 \qquad (2)$$

Next, we manipulated the degree to which they used others' knowledge. Specifically, we ran the utility value that performed memory retrieval based on individual knowledge to be 10 and the utility of productions that performed memory retrieval based on others' knowledge to be 8, 9, 10, 11, and 12. In figure 4, memory retrieval based on individual knowledge is "knowledge-retrieve-individual," and memory retrieval based on other's knowledge is "knowledge-retrieve-other". Probability is the probability that procedural knowl-

edge i will be used. Knowledge i is one of the currently selectable production rules. The equation for the selection rate of production is shown below, where U_i is the predicted utility, U_i(n) is the utility value for the nth time, α is the learning rate, and the reward received by the production on the nth application. We manipulated the parameters to simulate the effect of memory retrieval based on others' knowledge of knowledge retrieval and the use of text using the model.

$$Probability(i) = \frac{e^{U_i/\sqrt{2}s}}{\Sigma_j e^{U_j/\sqrt{2}s}} \qquad (3)$$

$$U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)] \qquad (4)$$

Result of Simulation

Reproducing Experimental Psychological Data This section reproduces the experimental psychological data to investigate the relationship between working memory and the number of correct and incorrect links. First, we investigated the relationship between working memory and the correct links. The average number of links was 0.5 for 0.7, 1 for 1.0, and 1.5 for 1.3. As working memory increases, memory is activated, and the number of correct links improves. Next, we checked whether working memory affected the reproduction of incorrect link counts. The average number of incorrect links was 1.5 for 0.7, 2 for 1.0, and 1.5 for 1.3, indicating that working memory did not affect the number of links, as in the experimental results. These results are consistent with those of the laboratory experiments.

The effect of reconstructing based on others Next, we compared the average number of correct links to explore the relationship between working memory and memory retrieval based on knowledge of others Figure 5 compares the number of links in memory retrieval based on working memory and the knowledge of others. The results show that the number of correct links in memory retrieval is based on others' knowledge when the working memory is low. When working memory is high, memory retrieval based on others' knowledge is better. Thus, This result indicates that the effect of reconstructing one's own knowledge based on others is only when working memory is high. Additionally, learners need to support working memory; the collaborative learning environment using concept maps reconsidering knowledge based on others' knowledge is essential.

The cognitive process of reconstructing based on others In this section, we examine in more detail why reconsidering one's knowledge based on others' knowledge increases the number of correct links. We focused on whether reconsidering one's knowledge based on others' knowledge enabled the correct reproduction of links in others' knowledge or whether the number of reproduced links increased simply because memory retrieval was performed more frequently.

First, we compared the number of reproductions of others' links to determine if we could reproduce the correct links
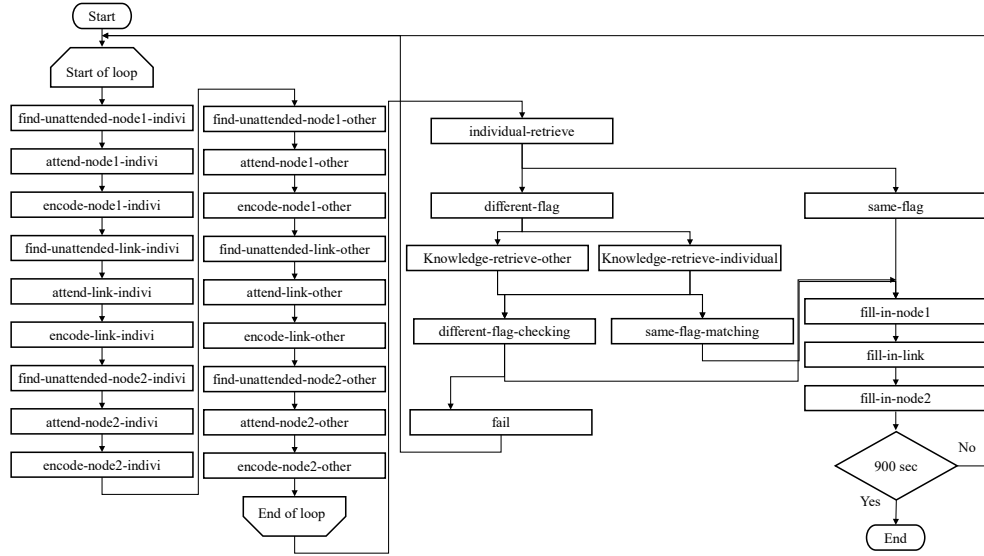
Figure 4: Flow of creating concept maps model
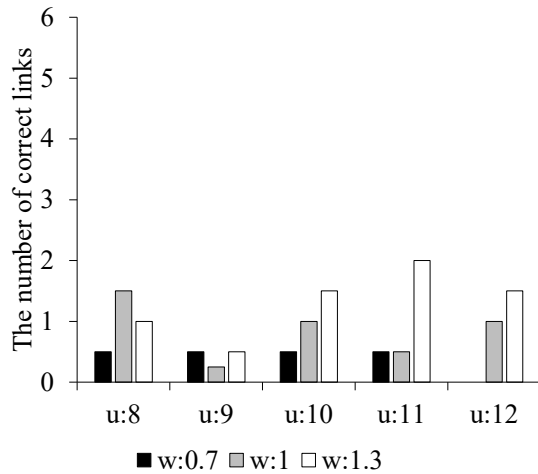
Figure 5: Comparison of correct links with working memory and reconstruction
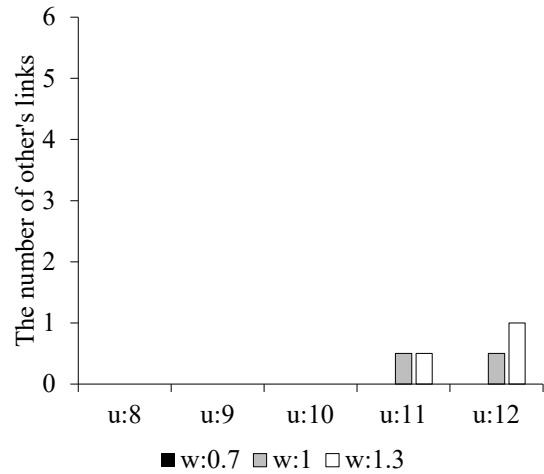
Figure 6: Comparison of the number of correct other's links with working memory and reconstruction based on others' knowledge

in other's knowledge. Figure 6 compares the number of reproductions of others' links in the working memory and the reconsidering of one's knowledge based on others' knowledge. We find that the number of links to others' knowledge increases as the rate of reconsidering one's knowledge based on others' knowledge increases. Therefore, the knowledge of others activated their inactivated declarative knowledge, and they could distribute their knowledge and select the proper knowledge among them. This tendency became slightly more substantial as the working memory increased.

Next, we compared the number of incorrect links to determine if the number of links increased simply because more memory retrieval was performed. Figure 7 compares the number of incorrect links in the reconsideration of one's knowledge based on working memory with other people's

knowledge. The results show that the number of incorrect links remains the same because the reconsideration of one's knowledge based on working memory and others' knowledge is greater. Therefore, it is clear that activation of declarative knowledge does not necessarily increase the number of incorrect concepts. The convergence of knowledge is not influenced by the reconsideration of one's knowledge or working memory but by the dispersion of knowledge. These results indicate that reconsidering one's knowledge based on others' knowledge activates declarative knowledge and enables correct knowledge. The cognitive processes of information divergence and convergence, in which knowledge is dispersed by reconsidering one's knowledge based on others' knowledge, are identified and selected from among them.
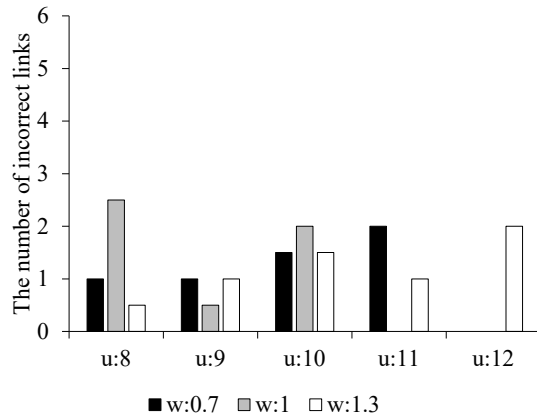
Figure 7: Comparison of the number of incorrect links with working memory and reconstruction based on others' knowledge.

Discussion

First, we examined replication in a psychological experiment and computational simulation about the relationships between working memory and correct and incorrect links. In a laboratory experiment, we found a correlation between working memory (2-back task) and the reproduction of correct knowledge and not a correlation between working memory and the reproduction of incorrect knowledge. Additionally, the simulation results with the working memory parameters were compatible with the data from the psychological experiment. Therefore, these results reveal that working memory influences collaborative learning.

Second, we examined the effect of reconstructing one's own knowledge based on others by observing the model's performance using parameters about working memory and the degree of reconsidering knowledge based on others. The results show that the number of correct links is not increased by reconstructing one's knowledge based on others' knowledge when the working memory is low. On the other hand, the number of correct links is increased by reconstructing one's knowledge based on others' knowledge when the working memory is high. The relationship between reconsidering one's knowledge based on others' knowledge and working memory suggests that support varies according to the level of working memory. When working memory is low, reconsidering one's knowledge based on others' knowledge is less critical. Therefore, working memory assistance is necessary and can explain the effectiveness of the visualization of others' knowledge (Engelmann & Hesse, 2010; Shimojo & Hayashi, 2019). When working memory is high, reconsidering knowledge based on others is effective. Facilitation to promote awareness of others' differences and reconsidering is necessary, which may explain the effectiveness of facilitating reconsidering based on others' knowledge (Hayashi, 2020).

Third, we examined the cognitive process of reconstructing one's own knowledge based on others by observing the model's performance in the view of other's links and incor-

rect links using parameters about working memory and the degree of reconsidering knowledge based on others. The results show that reconstructing one's knowledge based on others' knowledge increases the number of other links. On the other hand, reconstructing one's knowledge based on others' knowledge does not increase the number of incorrect links. Reconstructing one's knowledge based on others' knowledge does not simply increase reproduction and activation values. Specifically, the cognitive process of identifying and selecting proper knowledge from activated declarative knowledge by reconsidering one's knowledge based on others' knowledge is essential. The externalization of one's knowledge and that of others stores distributed knowledge in declarative knowledge, which is then elaborated and converged as task-relevant knowledge. The cognitive information processing model (Jorczak, 2011) also supports this result. The fact that misconceptions are not affected by working memory or the reconsideration of one's knowledge based on the knowledge of others indicates that knowledge dispersion is a more meaningful process. Collaboration effectiveness is related to knowledge divergence. The current results contribute to a better understanding of the cognitive process in collaborative learning because the cognitive model reproduces human cognitive processes. The limitation of this study is that collaborative learning was represented using a simple cognitive model of knowledge acquisition and recall, which could not reflect spoken discussions.

Conclusion

The results showed that working memory influenced collaborative learning using concept maps. The simulation results with the working memory parameters were compatible with the data from the psychological experiment. Additionally, the simulation indicated that reconsidering knowledge based on others is effective when working memory is high. Moreover, simulations regarding using others' knowledge and incorrect knowledge showed that reconsidering one's knowledge based on others' knowledge activated declarative knowledge and that identifying and selecting whether knowledge is correct were adequate for using correct knowledge. Externalizing one's and others' knowledge may have stored distributed knowledge in declarative knowledge, which was then elaborated and converged into task-relevant knowledge. The relationship between reconsidering one's knowledge based on others' knowledge and working memory suggests that support varies according to the level of working memory. In the future, it will be possible to develop a learning support system using the cognitive model of collaborative learning. By having an agent equipped with this cognitive model create a concept map and use it as a collaborative partner, we examined the effects of learning by teaching and being taught by a partner, peer tutoring.

References

Anderson, J., Corbett, A., Koedinger, L., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167–207.

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), (Vol. 2, p. 89-195). Academic Press.

Chi, M., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2020, 11). An Embedded-Processes Approach to Working Memory: How Is It Distinct From Other Approaches, and to What Ends? In *Working Memory: The state of the science.* Oxford University Press.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4*, 500–503.

Engelmann, T., & Hesse, F. (2010). How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning*, *5*, 299–319.

Hayashi, Y. (2020). Gaze awareness and metacognitive suggestions by a pedagogical conversational agent: an experimental investigation on interventions to support collaborative learning process and performance. *International Journal of Computer-Supported Collaborative Learning*, *15*(4), 469–498.

Hayashi, Y., & Koedinger, K. (2019). What are you talking about?: A cognitive task analysis of how specificity in communication facilitates shared perspective in a confusing collaboration task. In *Proceedings of the 41st annual conference of the cognitive science society* (pp. 1887–1893). Cognitive Science Society.

Hayashi, Y., & Shimojo, S. (2022a). Modeling perspective taking and knowledge use in collaborative explanation: Investigation by laboratory experiment and computer simulation using act-r. In *Proceedings of the 23rd international conference on artificial intelligence in education (aied2022)* (pp. 647–652).

Hayashi, Y., & Shimojo, S. (2022b). Relevant knowledge use during collaborative explanation activities: Investigation by laboratory experiment and computer simulation using act-r. In *Proceedings of 28th international conference on collaboration technologies and social computing (collabtech 2022)* (pp. 52–66).

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: towards a research agenda. *Education Tech Research*, *68*, 683–805.

Jorczak, R. L. (2011). An information processing perspective on divergence and convergence in collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *6*(2), 207–221.

Matsumuro, M., Miwa, K., Harada, E., Suto, S., Tomita, A.,

Makiguchi, M., & Jiajie, M. (2018). Aging effects on operations of an in-car device: Simulations focusing on time perception and activation spreading. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *25*(3), 279–292.

Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, *5*(6), 649–663.

Schweppe, J., & Rummer, R. (2014). Working memory, and long-term memory in multimedia learning: An integrated perspective based on process models of working memory. *Educational Psychology Review*, *26*, 285–306.

Shimojo, S., & Hayashi, Y. (2019). How shared concept mapping facilitates explanation activities in collaborative learning: an experimental investigation into learning performance in the context of different perspectives. In *Proceedings of the 27th international conference on computers in education (icce2019)* (pp. 172–177).

Shirouzu, H., & Miyake, N. (2002). Cognitively active externalization for situated reflection. *Cognitive Science*, *26*(4), 469–501.

Sjolund, L. A., Erdman, M., & Kelly, J. W. (2014). Collaborative inhibition in spatial memory retrieval. *Memory & Cognition*, *42*(6), 876–885.

Suebnukarn, S., & Haddawy, P. (2006). Modeling individual and collaborative problem-solving in medical problem-based learning. *User Modeling and User-Adapted Interaction*, *16*(3), 211–248.

Walker, E., Rummel, N., & Koedinger, K. (2014). Adaptive intelligent support to improve peer tutoring in algebra. *International Journal of Artificial Intelligence in Education*, *24*(1), 33–61.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). New York, NY: Springer New York.

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & education*, *46*(1), 71–95.

Modeling Fatigue in the N-Back Task with ACT-R and the Fatigue Module

Garrett Swan (gswan@aptima.com)
Aptima, Inc
Woburn, 01801 USA

Christopher A. Stevens (christopher.stevens.28@us.af.mil)
Air Force Research Laboratory
Wright Patterson AFB, OH USA

Bella Z. Veksler (b.veksler@tier1performance.com)
Tier1 Performance Solutions
Covington, KY, USA

Megan B. Morris (megan.morris.3@us.af.mil)
Air Force Research Laboratory
Wright Patterson AFB, OH USA

Abstract

Fatigue, if not managed properly, can have dangerous consequences for cognition and performance. It has been well established that fatigue impairs cognition, but theoretical development is necessary to better understand this relationship and predict conditions when performance may be at risk. In the present work, we examine a theory of fatigue situated in the ACT-R cognitive architecture. The theory proposes that fatigue results in the reduction of activation of task-relevant procedural and declarative knowledge. However, the relative impacts of fatigue on these two types of knowledge remains unclear. Here we investigated a task that requires activation of both procedural and declarative knowledge and we examined the fit of models assuming different fatigue mechanisms. Thirty-nine participants completed a 2-back task across 8 sessions over a 24-hour period. There was a significant effect of time on reaction time, hit rate, and false alarm rate. Our ACT-R variants of the N-back that included the fatigue module similarly showed an effect of time on those metrics. When comparing our variants to the behavioral data, the variant that included procedural lapses fit the data better than the variant that modeled fatigue as changes in activation strength and the variant that included both. These results provide information about the generalizability and boundary conditions of the mechanisms proposed by the ACT-R fatigue module.

Keywords: Fatigue, ACT-R, N-Back

Introduction

Human cognitive fatigue has costly implications in a variety of domains, such as aviation (Caldwell, 2005; Gaines, Morris, & Gunzelmann, 2020), railroad (Gertler, DiFiore, & Raslear, 2012), and medical (Kancherla et al., 2020) operations, among others. Fatigue is multifaceted and is affected by one's circadian and other biological rhythms (Achermann, 2004; Borbély & Achermann, 1992), sleep loss (Durmer & Dinges, 2005), and time-on-task (Doran, Van Dongen, & Dinges, 2001). Fatigue results in degradation to abilities such as reaction time, hand-eye coordination, situational awareness, and decision-making, and increases in risk-taking and errors of omission, among other effects to performance (Lamond & Dawson, 1999; Miller & Melfi, 2006), increasing safety risk. Modeling fatigue with cognitive models is useful because it provides quantitative predictions about how fatigue changes under different circumstances, such as variations in task or sleep schedules (Gunzelmann, Gross, Gluck, & Dinges, 2009), informing potential avoidance and mitigation strategies and tools.

Models of fatigue's impact on behavior have typically focused on the Psychomotor Vigilance Test (Dinges & Powell, 1985, PVT) and the development of a formalization of fatigue called the fatigue module (Gunzelmann, Gross, et al., 2009). That work has resulted in successful quantitative predictions of fatigue utilizing the Adaptive Control of Thought-Rational, ACT-R (Anderson et al., 2004), cognitive architecture. While other applications of the fatigue module have utilized more complex tasks than the PVT, such as a dual-task (Gunzelmann, Byrne, Gluck, & Moore Jr, 2009), lane-keeping (Gunzelmann, Moore Jr, Salvucci, & Gluck, 2011) and digit symbol substitution (Honn et al., 2020), less research has focused on the declarative memory aspect of ACT-R (e.g., see Gunzelmann, Gluck, Moore Jr, and Dinges 2012). We utilized ACT-R and the fatigue module in a simulation of the N-Back (Kirchner, 1958) to determine which mechanisms of the fatigue module generalize in a task that requires activation of both procedural and declarative knowledge.

Previous work has shown that the N-back is sensitive to sleep-deprivation (Choo, Lee, Venkatraman, Sheu, & Chee, 2005; Martínez-Cancino, Azpiroz-Leehan, & Jiménez-Angeles, 2015; Gerhardsson et al., 2019; Lythe, Williams, Anderson, Libri, & Mehta, 2012; Riontino & Cavallero, 2021). Here, instead of total sleep deprivation, we measured performance across 8 sessions in a 24-hour period while participants performed crew aviation tasks. In evaluating the fatigue module, we were interested in determining if changes in declarative activation alone could account for performance decrements as a function of fatigue, or if procedural lapses, as instantiated in the fatigue module, were necessary for an accurate simulation.

Methods

Participants

Forty-three pilots from Joint Base Charleston participated in the study; 39 had usable data for the current modeling effort ($M_{age} = 28$; $SD_{age} = 3.0$; Male = 32). The study was approved by the Air Force Research Laboratory Institutional Review Board.

Study Design

The study involved performing in a long-duration mobility simulator session with various mission tasks as a 3-person crew over a 24-hour period. During the 24 hours, participants underwent a cognitive battery at approximately 1200 (labeled Pre), 1315, 1650, 2326, 0230, 0456, 0925, and 1200 hours. Participants practiced the cognitive battery before the day of the simulator session. The cognitive battery consisted of the N-back, PVT, and Change Signal task (Brown & Braver, 2005;

Moore Jr & Gunzelmann, 2013). During breaks, participants were able to eat, drink, and take naps, with start and end time of the nap recorded by the researcher. The majority of the participants completed the cognitive battery at the specified times, though time points 0230 (77%) and 1200 (62%) had the lowest participation.

N-Back Task Description

Participants completed 150 trials of a 2-back task (Kirchner, 1958), where they were sequentially presented with a black letter (A, B, C, D, E, H, I, K, L, M, O, P, R, S, T) on a white display. Each trial was 4000 ms. The letter appeared on the screen for 500 ms at the start of the trial. When a letter appeared 2-back (e.g., in the sequence C-B-C, the first C is 2-back from the second C), participants responded by pressing the space bar on a keyboard. The probability of a 2-back on any given trial was set to 0.33.

ACT-R Model

The ACT-R cognitive architecture (Anderson et al., 2004) consists of discrete modules for distinct types of perceptual and cognitive processing (visual, audio, declarative memory, etc.). Buffers within the modules contain information about what is currently being perceived (visual module), what can be retrieved from declarative memory (declarative module), and how the model is interacting with a device (motor module). Information in the buffers is acted upon by productions, which are if-then statements based on information within the buffers that can be used to move information between different buffers or edit information within a buffer.

Fatigue module

The fatigue module was developed to add biologically plausible mechanisms of fatigue to ACT-R. The module utilizes information about sleep schedule to compute alertness (McCauley et al., 2013) that is then used to compute moment-by-moment fluctuations in ACT-R's procedural and/or declarative modules during a simulation. Whether the fatigue module affects either or both of the modules is up to the modeler.

When active for solely the procedural module, the fatigue module diminishes production utility as a function of variations in alertness due to circadian rhythm and sleep pressure. Additionally, response threshold is reduced (making errors of commission more likely) to simulate effort involved in staying awake and attentive. When no production fires, the model lapses through one cognitive cycle (50ms, plus noise) and the probability for another lapse to occur increases. Note that retrieval activation is tied to time in ACT-R, such that delays caused by lapses also impacts retrieval activation because declarative representations have more time to decay.

When active for solely the declarative module, alertness modulates global declarative activation, such that lower levels of alertness have lower activation values. Lower activation increases response time and makes it more likely that the the model will fail to retrieve task relevant information, resulting in misses in the 2-back task.

When active for both modules, declarative activation is modulated by both procedural lapses and the state of alertness.

N-Back ACT-R model

The model was designed to detect and respond to stimuli in the N-Back task environment. The model interacted with a custom built version of the N-Back in Python with the same task parameters experienced by the participants (parameters based on the psytoolkit parameters). ACT-R version 7.14 was used given that it is the most updated version of ACT-R compatible with the fatigue module.

The core model (Figure 1) was adapted from Held, Rieger, and Borst (2022). The model starts by building chunks in declarative memory that include the stimulus currently being attended and the previous stimulus. After the model has sufficient chunks in memory (2 chunks for a 2-back), the model performs a retrieval request to retrieve the previously shown letter (i.e., 1-back). If that retrieval is successful, the model then performs a second retrieval request to retrieve the letter preceding the previously shown letter (i.e., 2-back). The model compares the retrieved information to what is currently presented. If it is the same letter, then the model responds. Lastly, the model performs rehearsal (i.e., reactivating the chunk in declarative memory) by retrieving the current letter, and then retrieving the previous letter until the next letter appears.
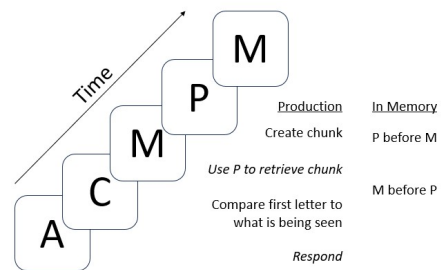
Figure 1: An illustration of the N-back task (text not to scale). The text to the right represents an abstraction of the model at the time of viewing the second M. Here, the model would utilize knowledge in the memory of the previous letter (P) to retrieve the chunk when P was presented. After a successful retrieval, the model then can compare the memory of the previous letter (M) to what is currently being presented (M) to determine if a response is necessary. We have italicized where errors could occur: retrieving the wrong chunk (influenced by parameter :mp), failing to retrieve a chunk at all (influenced by parameter :rt), and responding incorrectly (influenced by parameter :ppm).

Response Errors In the core model, errors could only occur by retrieving the wrong chunk given noise within the declarative memory system. The most likely outcome when retrieving the wrong chunk is the model not responding given that 14 out of 15 times it will not be the same letter. The ACT-R parameter most likely to cause an incorrect chunk to be retrieved is

:mp (i.e., the amount of a penalty applied when information in chunks mismatch), which was varied in the model fitting described below.

We added additional sources of error given that the core model did not consistently produce errors found in the behavioral data (Figure 2). Firstly, the model could fail to retrieve a chunk. Here, a failed retrieval would result in the model guessing. While true guessing would be responding on 0.33 trials, we found that 0.165 fit the data better given that participants had a low percentage of false alarms. The retrieval threshold (:rt) for declarative activation to be considered successfully retrieved was varied in model fitting.
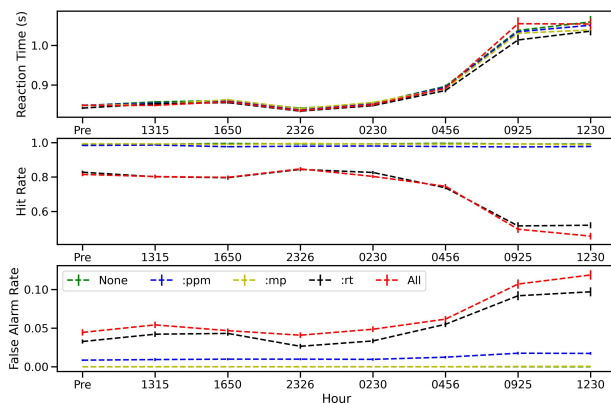
Figure 2: Reaction Time (Top), Hit Rate (Middle) and False Alarm Rate (Bottom) when varying ACT-R parameters. Here, we can visualize the effects of the different kinds of errors by viewing how :ppm, :rt, and :mp parameters individually affect performance. The "All" and "None" refer to versions with all 3 parameters turned on and off, respectively.

We also added motor noise that occasionally resulted in the model performing the wrong action to mirror how false starts occurred in the PVT ACT-R model (Gunzelmann, Gross, et al., 2009). The amount of motor noise was varied in model fitting. We set the similarity between responding and not responding to be 0.5 and let the :ppm (i.e., the partial production matching) ACT-R parameter vary.

Model Variants   The variants of the core model differed in which mechanism of the fatigue module was active. We examined 4 variants. (1) A baseline model with no fatigue module. To test whether performance changed with fatigue, one model had (2) declarative only active to test whether performance can be captured solely with a change to declarative memory activation and another model had (3) procedural only active to test whether performance can be captured with lapses. The final model was a combination of (4) procedural and declarative to test whether there was an interaction effect on retrieval activation.

Parameter Fitting   We performed parameter fitting such that each variant started at approximately the same level of performance and performance in the later parts of the session being a function of the different kinds of fatigue. Parameter fitting was achieved by comparing the model variants to participants' reaction time, hit rate, and false alarm rate on the first N-back completed before engaging in the simulation (labeled Pre). We performed a grid search on the following ACT-R parameters :mp, :rt, and :ppm from 2 to 3.2 with steps of 0.15 for each parameter, then a subsequent grid search around the best fitting combination with variations 0.1 above and below. The resulting values were the parameter fits (:mp, :rt, :ppm) for baseline (2.65, 3.1, and 2.4), declarative only (2.7, 3.2, 2.7), procedural only (2.9, 2.9, 3.05), and combined (2.65, 2.9, 2.9) variants. All other ACT-R parameters were kept the same between the different variants.

Fatigue module parameter [1] values were selected based on existing fatigue module models (e.g., Walsh, Gunzelmann, and Van Dongen 2017) and thus in line with typical values in lieu of a parameter search. Specifically, when the procedural fatigue was active, we set initial utility to 2.6, the initial threshold to 2.2, :fpbmc to 0.02, :utbmc to 0.01, and reset :fp-dec to 0.9875 each time a stimulus appeared. When declarative fatigue was active, we set :fdbmc to -0.015, :fdc to 0.97, and reset fd:dec to 0.9925 each time a stimulus appeared. The only parameter that was varied for the combined variant was changing :fp-dec to 0.9925 to mitigate performance being drastically worse in the combined variant.

Performance Measures

Trial Simulation   We simulated the same number of participants, trials, and stimulus parameters as the original study for each model variant. Given the impact of sleep schedule on the fatigue module, we used self-reported information about nap times that was collected during the original study. We approximated those schedules as input into the fatigue module, resulting in 4 schedules approximately equally divided in the simulations: (1) one hour naps at 1500, 2230, and 0900, (2) a two hour nap at 1730 and a one hour nap at 0100, (3) a two hour nap at 2200 and a one hour nap at 0400, and (4) a three hour nap at 1700.

Dependent Measures   Our dependent measures of interest were average reaction time (RT) on correct trials, hit rate (correct response / (correct response + misses)), and false alarm rate (false alarm / (false alarm + correct rejection). To determine the effects of fatigue, we used linear mixed effects models (statsmodels mixedlm, Seabold and Perktold 2010), with participant as a random factor.

To determine which model variant accounts for the data best, we took the mean of the model's performance for each dependent measure, then calculated root mean squared error

---

[1]The fatigue module parameters we altered were the fp and fd biomath constants (:fpbmc and :fdbmc), the fd constant (:fdc), the utility threshold biomath constant (:utbmc), and fp-dec and fd-dec that reduces fp-percent and fd-percent each time there is a lapse.

(RMSE) then normalized the RMSE (NRMSE). We interpreted the model with the lowest average NRMSE across the dependent measures as the best fitting model.

Results

Reaction Time

We found a significant effect of time on reaction time for the behavioral data [$b = 0.005$, $se = 0.001$, $z = 5.9$, $p < 0.001$], the declarative variant [$b = 0.001$, $se < 0.001$, $z = 15.2$, $p < 0.001$], the procedural variant [$b = 0.008$, $se < 0.001$, $z = 16.4$, $p < 0.001$], and the combined variant [$b = 0.008$, $se < 0.001$, $z = 16$, $p < 0.001$]. Reaction time increased with time (Figure 3). As expected, reaction time did not significantly change as a function of time in the baseline variant [$b < -0.001$, $se < 0.001$, $z = -1.56$, $p = 0.12$].
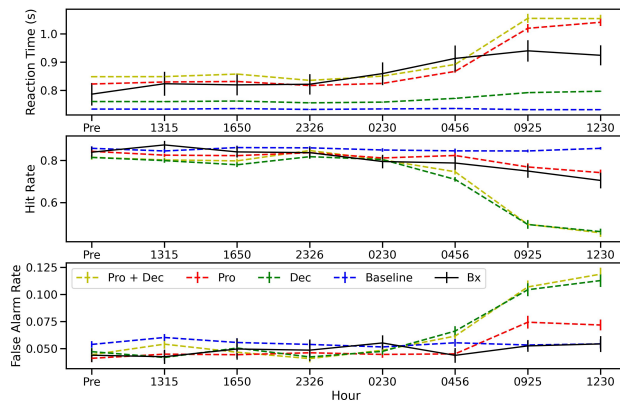

Figure 3: Reaction Time (Top), Hit Rate (Middle), and False Alarm Rate (Bottom) for the behavioral data (Bx), baseline variant (Baseline), the variant with declarative fatigue (Dec), the variant with procedural fatigue (Pro), and the variant with combined procedural and declarative (Pro + Dec).

Hit Rate

We found a significant effect of time on hit rate for the behavioral data [$b = -0.005$, $se = 0.001$, $z = 6.05$, $p < 0.001$], the declarative variant [$b = -0.012$, $se = 0.001$, $z = -16.9$, $p < 0.001$], the procedural variant [$b = -0.003$, $se < 0.001$, $z = -6.5$, $p < 0.001$], and the combined variant [$b = -0.012$, $se = 0.001$, $z = -15.3$, $p < 0.001$]. Hit rate decreased with time (Figure 3). As expected, hit rate did not significantly change as a function of time in the baseline variant [$b < -0.001$, $se < 0.001$, $z = -0.46$, $p = 0.65$].

False Alarm Rate

We found a significant effect of time on false alarm rate for the behavioral data [$b = 0.001$, $se = 0.001$, $z = 2.06$, $p < 0.05$], the declarative variant [$b = 0.002$, $se < 0.001$, $z = 12.6$, $p < 0.001$], the procedural variant [$b = 0.001$, $se < 0.001$, $z = 6.6$, $p <$

0.001], and the combined variant [$b = 0.002$, $se < 0.001$, $z = 11.7$, $p < 0.001$]. False alarm rate decreased with time (Figure 3). As expected, false alarm rate did not significantly change as a function of time in the baseline variant [$b < -0.001$, $se < 0.001$, $z = -0.84$, $p = 0.4$].

Evaluating model variants

The RMSEs for RT were 0.278, 0.262, 0.249, and 0.252 for the baseline, declarative, procedural, and combined models. The procedural and combined variants fit reaction time best, while the declarative variant's RT increased much more slowly than the behavioral data. For hit rate, the RMSEs were 0.188, 0.212, 0.179, and 0.21 for the baseline, declarative, procedural, and combined models. The declarative and combined both overestimated the impact of fatigue on accuracy, while the procedural model accounted for the data best. The RMSEs for false alarm rate were 0.047, 0.054, 0.047, and 0.054 for the baseline, declarative, procedural, and combined models.

To compare across dependent measures, we then normalized the RMSEs by dividing RMSE by the average value from the behavioral data. The averaged NRMSEs for the variants was 0.5035, 0.5556, 0.4921, 0.5564 for the baseline, declarative, procedural, and combined models. The procedural model was 12.9% lower than the next best NRMSE from a variant that showed an effect of fatigue on performance (i.e., the declarative variant), indicating fairly strong evidence in favor of the procedural model.

Discussion

We replicated existing work demonstrating an effect of fatigue on performance in the *N*-Back task (Choo et al., 2005; Martínez-Cancino et al., 2015; Gerhardsson et al., 2019; Lythe et al., 2012; Riontino & Cavallero, 2021). We implemented a model in ACT-R and evaluated different fatigue mechanisms using the fatigue module to determine which mechanisms best accounted for changes in behavior as a function of fatigue. Overall, the results suggest procedural lapses were sufficient to account for performance changes as a function of fatigue in the *N*-Back.

The participants and all of the model variants with the fatigue module showed an effect of fatigue on reaction time, hit rate, and false alarm, indicating strong support for needing some mechanism of fatigue when measuring performance across time. The variants differed on the magnitude of the effect of fatigue on performance. The procedural variant fit best, though the declarative and combined variant still captured qualitative changes in performance across time. One factor that could improve the fits of the models is the inclusion of individual differences (e.g., see Fisher, Morris, Stevens, and Swan 2024 for potential pitfalls with one-size-fits-all models). Previous research has shown that individuals experience neurobehavioral deficits of sleep differently (Van Dongen, Baynard, Maislin, & Dinges, 2004), which could be instantiated by allowing fatigue module parameters to vary across simulations.

The variant based off of declarative activation had limited flexibility in our demonstration, given that activation value determined reaction time (i.e., the lower the activation, the slower the retrieval). Significantly changing activation strength to create reaction times more in line with behavior would also impact the model's hit and false alarm rate. The implication is that participants were experiencing attentional lapses in the *N*-Back which accounted for the reaction time differences. In the PVT, lapses are typically identified by looking at responses above 500 ms, which is in the tail of the distribution in the PVT task. Altering the design of the *N*-Back by increasing the inter-stimulus interval or by utilizing Electroencephalography (EEG) metrics (Curley, Borghetti, & Morris, 2024) could more explicitly measure lapses in the *N*-Back.

All of the models tended to overestimate the effects of fatigue during the last two measurements (hours 0925 and 1230). In other words, while participants still performed worse during those sessions of the *N*-Back, the models predicted even worse performance. The difference was likely not driven by difference in circadian phase, given that the fatigue module accounts for those fluctuations. Individuals may have benefited from caffeine use during break periods (Halverson, Myers, Gearhart, Linakis, & Gunzelmann, 2022). Another factor that could have been driving better than expected performance was a macro-level end-spurt effect (Morris, Haubert, & Gunzelmann, 2020). In this study, participants knew when the experiment was going to end, thus they have been more engaged towards the end of the study. However, end-spurt effects have typically been explored within session, not across multiple time points, so future research may be necessary to determine if an end-spurt effect could occur across measurement periods.

One of the goals of this comparison was to determine if modeling the *N*-Back could provide additional constraints, or support of, the declarative memory fatigue mechanism in the fatigue module. Interestingly, that was not the case, given that the combined variant produced fits worse than the procedural variant alone. This could simply reflect the *N*-Back task's poor relationship with existing measures of memory (e.g., Owen, McMillan, Laird, and Bullmore 2005), with some suggesting that recognition in the *N*-Back may better reflect attentional control (Kane, Conway, Miura, & Colflesh, 2007). Perhaps investigating other paradigms with similar comparisons of model variants could be useful. For example, in a study utilizing change detection, participants were found to have attentional lapses that directly impacted working memory performance (DeBettencourt, Keene, Awh, & Vogel, 2019; Adam, Mance, Fukuda, & Vogel, 2015). In such a task, one could model if procedural lapses alone could account for the changes in working memory performance or if the declarative components of the fatigue module provide better fits to the data.

Our findings provide additional support for the generalizability of the fatigue module in providing quantitative predictions about the effects of fatigue. Despite the complexity difference in terms of ACT-R productions between the *N*-Back

(here, 19) and PVT (3), the fatigue module captured both quantitative changes of performance as a function of fatigue. Future research involving the fatigue module should continue testing the scalability in more complex tasks.

Acknowledgments

References

Achermann, P. (2004). The two-process model of sleep regulation revisited. *Aviation, Space, and Environmental Medicine*, *75*(3), A37–A43.

Adam, K. C., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience*, *27*(8), 1601–1616.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036.

Borbély, A. A., & Achermann, P. (1992). Concepts and models of sleep regulation: an overview. *Journal of Sleep Research*, *1*(2), 63–79.

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307*(5712), 1118–1121.

Caldwell, J. A. (2005). Fatigue in aviation. *Travel Medicine and Infectious Disease*, *3*(2), 85–96.

Choo, W.-C., Lee, W.-W., Venkatraman, V., Sheu, F.-S., & Chee, M. W. (2005). Dissociation of cortical regions modulated by both working memory load and sleep deprivation and by sleep deprivation alone. *Neuroimage*, *25*(2), 579–587.

Curley, T. M., Borghetti, L., & Morris, M. B. (2024). Gamma power as an index of sustained attention in simulated vigilance tasks. *Topics in Cognitive Science*, *16*(1), 113–128.

DeBettencourt, M. T., Keene, P. A., Awh, E., & Vogel, E. K. (2019). Real-time triggering reveals concurrent lapses of attention and working memory. *Nature Human Behaviour*, *3*(8), 808–816.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual rt task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, *17*(6), 652–655.

Doran, S. M., Van Dongen, H. P., & Dinges, D. F. (2001). Sustained attention performance during sleep deprivation: evidence of state instability. *Archives Italiennes de Biologie*, *139*(3), 253–267.

Durmer, J. S., & Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. In *Seminars in neurology* (Vol. 25, pp. 117–129).

Fisher, C. R., Morris, M. B., Stevens, C. A., & Swan, G. (2024). The role of individual differences in human-automated vehicle interaction. *International Journal of Human-Computer Studies*, *185*, 103225.

Gaines, A. R., Morris, M. B., & Gunzelmann, G. (2020). Fatigue-related aviation mishaps. *Aerospace Medicine and Human Performance*, *91*(5), 440–447.

Gerhardsson, A., Åkerstedt, T., Axelsson, J., Fischer, H., Lekander, M., & Schwarz, J. (2019). Effect of sleep deprivation on emotional working memory. *Journal of Sleep Research*, *28*(1), e12744.

Gertler, J., DiFiore, A., & Raslear, T. (2012). Fatigue status of the us railroad industry (tech. rep. dot/fra/ord-13/06). *US Department of Transportation, Washington, DC: Federal Railroad Administration*.

Gunzelmann, G., Byrne, M. D., Gluck, K. A., & Moore Jr, L. R. (2009). Using computational cognitive modeling to predict dual-task performance with sleep deprivation. *Human Factors*, *51*(2), 251–260.

Gunzelmann, G., Gluck, K. A., Moore Jr, L. R., & Dinges, D. F. (2012). Diminished access to declarative knowledge with sleep deprivation. *Cognitive Systems Research*, *13*(1), 1–11.

Gunzelmann, G., Gross, J. B., Gluck, K. A., & Dinges, D. F. (2009). Sleep deprivation and sustained attention performance: Integrating mathematical and cognitive modeling. *Cognitive Science*, *33*(5), 880–910.

Gunzelmann, G., Moore Jr, L. R., Salvucci, D. D., & Gluck, K. A. (2011). Sleep loss and driver performance: Quantitative predictions with zero free parameters. *Cognitive Systems Research*, *12*(2), 154–163.

Halverson, T., Myers, C. W., Gearhart, J. M., Linakis, M. W., & Gunzelmann, G. (2022). Physiocognitive modeling: Explaining the effects of caffeine on fatigue. *Topics in Cognitive Science*, *14*(4), 860–872.

Held, M., Rieger, J. W., & Borst, J. P. (2022). Multitasking while driving: central bottleneck or problem state interference? *Human Factors*, 001872082221143857.

Honn, K. A., Halverson, T., Jackson, M., Krusmark, M., Chavali, V., Gunzelmann, G., & Van Dongen, H. (2020). New insights into the cognitive effects of sleep deprivation by decomposition of a cognitive throughput task. *Sleep*, *43*(7), zsz319.

Kancherla, B. S., Upender, R., Collen, J. F., Rishi, M. A., Sullivan, S. S., Ahmed, O., . . . others (2020). Sleep, fatigue and burnout among physicians: an american academy of sleep medicine position statement. *Journal of Clinical Sleep Medicine*, *16*(5), 803–805.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the n-back task: a question of construct validity. *Journal of Experimental psychology: Learning, Memory, and Cognition*, *33*(3),

615.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352.

Lamond, N., & Dawson, D. (1999). Quantifying the performance impairment associated with fatigue. *Journal of Sleep Research*, *8*(4), 255–262.

Lythe, K. E., Williams, S. C., Anderson, C., Libri, V., & Mehta, M. A. (2012). Frontal and parietal activity after sleep deprivation is dependent on task difficulty and can be predicted by the fmri response after normal sleep. *Behavioural Brain Research*, *233*(1), 62–70.

Martínez-Cancino, D., Azpiroz-Leehan, J., & Jiménez-Angeles, L. (2015). The effects of sleep deprivation in working memory using the n-back task. In *Vi latin american congress on biomedical engineering claib 2014, paraná, argentina 29, 30 & 31 october 2014* (pp. 421–424).

McCauley, P., Kalachev, L. V., Mollicone, D. J., Banks, S., Dinges, D. F., & Van Dongen, H. P. (2013). Dynamic circadian modulation in a biomathematical model for the effects of sleep and sleep loss on waking neurobehavioral performance. *Sleep*, *36*(12), 1987–1997.

Miller, J. C., & Melfi, M. L. (2006). Causes and effects of fatigue in experienced military aircrew. *Brooks City-Base, USA: Air Force Research Laboratory*.

Moore Jr, L. R., & Gunzelmann, G. (2013). Task artifacts and strategic adaptation in the change signal task. *Cognitive Systems Research*, *24*, 35–42.

Morris, M. B., Haubert, A. R., & Gunzelmann, G. (2020). Beyond the vigilance end-spurt with event-related potentials. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 64, pp. 1258–1262).

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*(1), 46–59.

Riontino, L., & Cavallero, C. (2021). Individual differences in working memory efficiency modulate proactive interference after sleep deprivation. *Psychological Research*, *85*(2), 480–490.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th python in science conference* (Vol. 57, pp. 10–25080).

Van Dongen, P., Baynard, M. D., Maislin, G., & Dinges, D. F. (2004). Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. *Sleep*, *27*(3), 423–433.

Walsh, M. M., Gunzelmann, G., & Van Dongen, H. P. (2017). Computational cognitive modeling of the temporal dynamics of fatigue from sleep loss. *Psychonomic Bulletin & Review*, *24*, 1785–1807.

Rational Compression in Choice Prediction

Max Taylor-Davies (s2227283@ed.ac.uk)
School of Informatics
University of Edinburgh

Christopher G. Lucas
School of Informatics
University of Edinburgh

Abstract

To successfully navigate its social environment, an agent must construct and maintain representations of the other agents that it encounters. Such representations are useful for many tasks, but they are not without cost. As a result, agents must make decisions regarding how much information they choose to store about the other agents in their environment. Using choice prediction as an example task, we illustrate the problem of finding agent representations that optimally trade off between downstream utility and information cost, before presenting the results of two behavioural experiments designed to examine this tradeoff in human social cognition. We find that people are sensitive to the balance between representation cost and downstream value, while still deviating from optimality.

Keywords: social cognition, resource rationality, decision-making, information theory

Introduction

In order to produce adaptive behaviour, an agent must acquire and maintain an internal representation of its environment (Craik, 1943; Tolman, 1948; Wilson et al., 2014). For instance, a foraging animal should have some representation of which areas of its environment are most likely to provide food, as well as which might contain sources of danger to avoid. But this is true not only for the inanimate features of the world— unless condemned to an entirely solitary existence, we can expect that many environments encountered by a hypothetical agent will contain *other agents*. Much as an agent should represent the rest of the environment, we expect that it ought also to represent these other agents. Humans do this, of course—in fact, it seems we automatically form mental representations of the other people we encounter (Dennett, 1987; Malle, 2008; Baker et al., 2017). We use these representations for a variety of different purposes: understanding the strengths and weaknesses of a colleague to effectively collaborate with them; determining whether a stranger should be treated as friend or foe; or predicting the plays of a chess opponent in order to defeat them. In general, a detailed representation of the world is more useful than a coarse one, in the sense of allowing greater predictive power or insight. But real agents, whether biological or artificial, inevitably have to contend with limits on their cognitive or computational resources. We therefore do not typically expect an agent to hold within their mind a 1:1 lossless model of the world; instead, they will employ a representational system that involves some degree of approximation or compression. Indeed, the argument for compression is perhaps especially clear in the specific case of representing other agents. As soon as we allow for the fact that this process

goes both ways (i.e. as I represent agent X, agent X in turn represents me) then we have to contend with some level of recursion: my representation of agent X must contain within it some representation of myself. For these representations to involve no loss of information, my mind would have to contain within it a number of perfect copies of itself, which cannot be possible. This line of thinking motivates us to consider two related questions. First, how much information *should* an optimal agent represent about the other agents in its environment? And second, is the answer to this question reflected in the choices that people actually *do* make in response to this problem?

Over recent years, there has been a growing body of work in cognitive science that seeks to understand human cognition through the lens of *resource rationality* (Lieder & Griffiths, 2019; Bhui et al., 2021; Icard, 2023). As a framework, resource rationality extends the classic ideas of decision theory (Neumann & Morgenstern, 1953; Jeffrey, 1965) and rational analysis (Anderson, 1990) to account for the notion that agents do not possess infinite capacity for acquiring, storing or processing information. It can also be seen as building on the concept of *bounded* rationality popularised by Simon (among others), while being more explicit in its focus on the idea of *resourcefulness*, i.e. of agents making the most effective use of the cognitive resources available to them. Various formalisations of this idea are possible (Icard, 2023); we will adopt a version of what Icard terms the 'cost-theoretic approach', which considers a continuous tradeoff between the *utility* of a given cognitive or behavioural strategy and the cost of carrying it out. Note that this still leaves considerable flexibility via the choice of how both sides of this tradeoff are defined. As far as cost is concerned, our focus in this paper is specifically on information cost; i.e. the cost of acquiring and storing the representations (of other agents) that support a particular strategy. This is distinct from the computational cost of converting those representations into decisions or behaviour. While a complete analysis should account for both, we leave this for future work, and will focus in this paper on a task setting in which the optimal decision strategy is extremely simple given an appropriate representation.

Task

General objective

In its most general form, the cost-theoretic approach to resource rationality is concerned with maximising an objective

function that looks like this:

$$R := S - \lambda C \qquad (1)$$

where S is some measure success or performance on our task of interest, C is some measure of the cost(s) we want to minimise, and λ is a tradeoff parameter that governs the relative weight assigned to each quantity. For our purposes, we make R a function of some chosen social representation χ:

$$R(\chi) := S(\chi) - \lambda C(\chi). \qquad (2)$$

For any choice of (S, C, λ), the optimal representation is then given by $\chi^* = \arg\max_\chi R(\chi)$. This optimality criterion is similar to the objectives used within work on capacity-limited Bayesian decision-making and RL, such as Arumugam et al. (2024). The key difference (beyond our explicit focus on social representations) is that we are interested not so much in the cognitive cost of converting representations into behaviour, but in the cost the representations themselves. In general, we expect this be a combination of the cost involved in acquiring a representation (i.e. inferring it from observation), and the cost involved in storing it—for now we adopt a simplistic definition of $C(\chi)$ as the *number of bits* in χ, assuming that representations which require a greater number of bits to store or transmit will impose a higher cognitive cost.

Pairwise choice prediction

As for S, we construct a minimal social cognition task where one agent (Alice) tries to predict the choices made by a second agent (Bob). First, let S be some choice space. S can in general contain any sort of thing that an agent could make choices over; we will say here that it is the space of possible states of the environment. At trial t, we sample a random pair of states (s_1, s_2) uniformly from S, and Alice makes a prediction c_{pred} about which state Bob will choose. Bob then makes his choice c_{actual}—if Alice's prediction was correct ($c_{\text{pred}} = c_{\text{actual}}$), she earns a reward. Alice's goal is to maximise her total reward earned over some large number of trials. This task is attractive in its conceptual simplicity—but it does also bear a relation to more realistic problems faced by people navigating social environments, such as predicting the lane choice of other drivers on the road, or which of two possible gifts your partner would prefer.

Of course, how well Alice can in principle do on this task depends on how Bob makes his choices. We will assume that Bob is a noisily rational agent whose decisions are described by a Boltzmann choice rule:

$$\Pr[\text{choose } s_1] = \frac{\exp\left(\frac{u(s_1)}{\beta}\right)}{\exp\left(\frac{u(s_1)}{\beta}\right) + \exp\left(\frac{u(s_2)}{\beta}\right)} \qquad (3)$$

where $u : S \to \mathbb{R}$ is Bob's utility function, which maps elements of S to scalar utilities, and β quantifies his 'decision noise' (i.e. the extent to which he deviates from optimal choice behaviour). Given this, and assuming access to some approximate representation $\hat{u}$ of Bob's true utility function (defined

over the same state space), the optimal strategy is clearly to make predictions as

$$c_{\text{pred}} \mid \hat{u}, (s_1, s_2) = \arg\max_{s \in (s_1, s_2)} \hat{u}(s) \qquad (4)$$

Using a 0-1 loss, the objective function for a single trial is given by

$$R_{\text{trial}}(\hat{u}) := \mathbb{1}(c_{\text{pred}} \mid \hat{u} = c_{\text{actual}}) - \lambda n_{\text{bits}}(\hat{u}) \qquad (5)$$

To obtain the general objective function, over both trials and different instances of Bob (with different u), we will treat $\hat{u}$ as a random variable resulting from the application of some 'representation scheme' to the true utility function u. We can then take the expectation over both state-pairs and $\hat{u}$ to write

$$R_{\text{expected}}(\hat{u}) := \mathbb{E}_{S^2}[\mathbb{1}(c_{\text{pred}} \mid \hat{u} = c_{\text{actual}})] - \lambda H[\hat{u}] \qquad (6)$$

where H denotes the differential entropy.

If we only have to represent a very small number of agents, or a small state space S, then it may be feasible to represent utility functions exactly (i.e. use $\hat{u} = u$), even for $\lambda > 0$. But if the agent population or state space is large, or if $\lambda \gg 0$, then the optimal representation in terms of Equation 6 will likely be an approximation $\hat{u}$ that discards some information for the sake of lower entropy. A nice consequence of the simplicity of our prediction task is that we can write out an analytical expression for the expected success (i.e. prediction accuracy) given an arbitrary $\hat{u}$:

$$\mathbb{E}_{S^2}\left[\mathbb{1}(c_{\text{pred}} \mid \hat{u} = c_{\text{actual}})\right] = \frac{1}{2} + \frac{1}{2} \mathbb{E}_{S^2}\left[\text{sign}(\Delta u \, \Delta \hat{u}) \tanh\left(\frac{\Delta u}{2\beta}\right)\right] \qquad (7)$$

where $\Delta u = u(s_1) - u(s_2)$ and $\Delta \hat{u} = \hat{u}(s_1) - \hat{u}(s_2)$. A derivation for this expression is given in Appendix B, but the intuition here is that the prediction accuracy given $\hat{u}$, relative to the prediction accuracy given u, depends on the probability that $\hat{u}$ can correctly resolve the 'polarity' of a pair of states resolved by u. The objective function in Equation 6 can then be written as

$$R_{\text{expected}}(\hat{u}) = \frac{1}{2} + \frac{1}{2} \mathbb{E}_{S^2}\left[\text{sign}(\Delta u \, \Delta \hat{u}) \tanh\left(\frac{\Delta u}{2\beta}\right)\right] - \lambda H[\hat{u}] \qquad (8)$$

Compression through state aggregation

So far, we have just considered the idea of approximate representations in the abstract. But what might these approximate representations actually look like? One straightforward way to approximate a utility function is through state aggregation—i.e. group all states within a given-sized 'patch' of S under a single value (Sutton & Barto, 2018; Abel et al., 2019). It is important here to note that we do not take this to be an optimal (or even particularly strong) compression strategy for any given state space S—but its simplicity and generality makes it an attractive choice for illustrating the tradeoff dynamics that we are concerned with. To do this, we set up a simulation
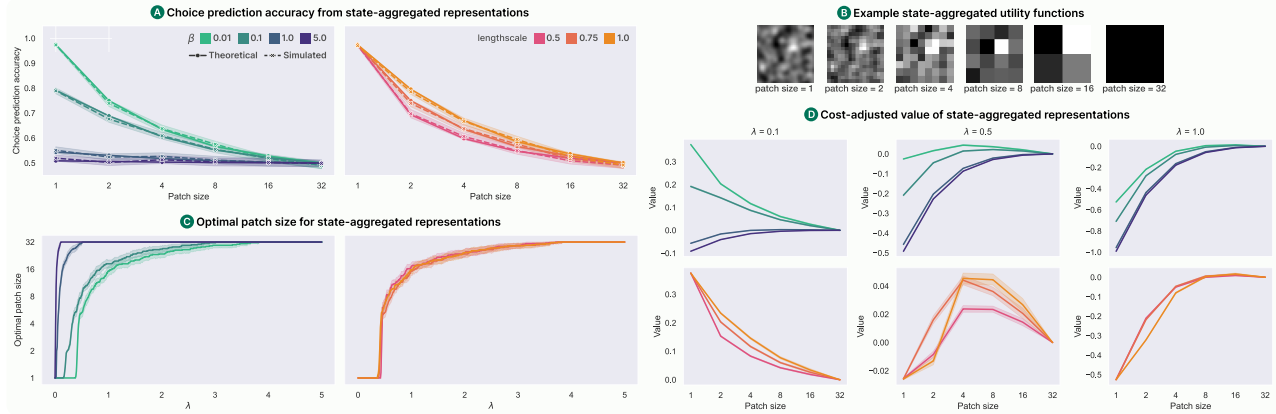
Figure 1: Results of simulating pairwise choice prediction with state-aggregated utility function representations, for noisily rational target agents with spatially correlated 2D utility functions. **(A)** theoretical (Equation 7) vs simulated prediction accuracy as a function of aggregation patch size for different values of utility function lengthscale and target agent β. **(B)** illustration of state aggregation levels for example 2D utility function. **(C)** optimal patch size as a function of increasing tradeoff parameter λ. **(D)** cost-adjusted return (Equation 8) as a function of patch size, with cost given by continuous entropy.

environment where agents make choices over pairs of tiles in a 32x32 2D grid. Each agent has a spatially correlated utility function drawn from a Gaussian Process (with RBF kernel), and a decision noise parameter β. Over a number of trials, we simulate a choice prediction strategy using different levels of state aggregation (where state aggregation is measured by 'patch size', i.e. the number of grid tiles grouped under a single value in the aggregated representation). For this simple setup, the specific relationship between patch size and prediction accuracy should be determined by both the lengthscale of the Gaussian Process (i.e. how smoothly u varies over $\mathcal{S}$), and the decision noise β of the agents making the choices—we therefore repeat the simulation for different values of each parameter (keeping the other constant).

The results of these simulations are shown in Figure 1. First, panel **(A)** shows that Equation 7 successfully captures the effect on simulated choice prediction accuracy of increasing patch size, across all simulated values of lengthscale and β. As we would expect, prediction accuracy decreases monotonically with increasing patch size. Furthermore, for any given patch size < 32, prediction accuracy decreases with increasing β (i.e. as agents become more unpredictable). We also see that the decrease in prediction accuracy with patch size is less steep at higher lengthscale (i.e. smoother u), where less information is lost for a given amount of aggregation. Panel **(D)** illustrates the expected cost-adjusted return (Equation 8) as a function of patch size, for the same set of β and lengthscale values, and for various values of the tradeoff parameter λ. We see that the optimal aggregation level is shifted to the right as we increase λ (and thus care more about information cost). This same trend is also seen in panel **(C)**, which shows directly how the optimal patch size changes as a function of λ.

Figure 2: Interface for behavioural experiments

Experiment 1

In the preceding sections, we presented a theoretical and computational analysis of the tradeoff between information cost and predictive value faced by agents representing others' utility functions. We now seek to shed light on this tradeoff in *human* social cognition—that is, does our optimal analysis predict people's actual choices about how much information to represent about other agents' utility functions?

Procedure

To answer this question, we developed a behavioural experiment based on the simple pairwise choice prediction task outlined above. We recruited a total of $n = 90$ adults through the online platform Prolific, who were then directed to an on-

line game consisting of 5 rounds (excluding an initial tutorial round), and instructed to try and maximise their final score. To incentivise performance, participants were rewarded with bonus payments for achieving scores above a certain threshold. The cover story for the task was that participants (playing the role of Alice) had to predict the choices made between different plots of land (tiles) in a field (8x8 2D grid) by a farmer (playing the role of Bob) trying to grow a particular crop. While making choice predictions, participants were given access to a (possibly aggregated) representation of the farmer's utility function in the form of a map of the different plots' 'quality' for growing the crop in question. Crucially, the tradeoff dynamics were introduced via a game mechanic were participants *spent* points to acquire this map, and *earned* points for correct predictions. At the start of each round, participants selected a level of state aggregation, paying a cost in points determined by the level of aggregation chosen. They were then presented with a series of randomly sampled tile pairs—for each pair they were instructed to select the tile that they thought agent would choose, while being able to see the (possibly aggregated) map they had just 'purchased'. For simplicity, the farmer was set to be perfectly rational, i.e. $\beta \to 0$ under the choice rule in Equation 3. Figure 2 shows the main components of the game interface. We varied two factors between participants in a 2x3 design (with 15 participants per condition): the texture of the 2D utility functions ('rough', 'smooth', corresponding to GP lengthscales of 0.5 and 2 respectively), and the absolute costs of the different maps ('low', 'medium', 'high'). All participants faced the same number and sequencing of rounds, regardless of condition, and the *relative* cost of the different maps was always the same. To maximise their overall score, a given participant would need to choose, at each round, the level of state aggregation that optimally balanced cost against expected predictive value (depending on their assigned condition). We recorded participants' choices of aggregation level at each round, as well as all of the pairwise choice predictions that they made.

Results

From our behavioural data, we compute participants' average prediction accuracy as a function of aggregation level (patch size), split by texture condition. We then compare these in Fig 3 (A) to Equation 7. We can see that participants' average prediction accuracy is fairly well captured by the model—that is, participants in general made effective use of the information contained in their chosen representations. Participants were also more accurate in the smooth utility function condition, reflecting the fact that less information is lost when aggregating spatially correlated functions with higher lengthscale. So, our model predicts how participants' prediction accuracy varies with aggregation level. But can it predict which aggregation levels participants will select? For each of the 6 conditions, we compare the recorded proportions of participants' patch size selections against the choice distribution given by three

different variants of a noisily rational model

$$\Pr\{\text{select } \hat{u}\} \propto \exp\left(\frac{V_m(\hat{u})}{\beta}\right) \quad (9)$$

where $V_m(\hat{u})$ is set as either the expected accuracy, the negative cost, or the full cost-adjusted return (from Equation 8). This comparison is shown in Figure 3 (B), using $\beta = 0.25$. While none of these three models is able to capture participants' patch size selections perfectly, it is clear that the full resource-rational choice rule is a much better fit than either the accuracy-only or cost-only models—indicating that to at least some extent, participants are sensitive to the tradeoff between information cost and predictive value in selecting representations. For instance, participants' selection probability decreased monotonically with increasing patch size in the low-cost condition, and increased almost monotonically in the high-cost condition. However, for the medium cost condition, the resource-rational model predicts a greater difference in selection probabilities between the rough and smooth conditions than was reflected in participants' behaviour. This suggests that participants in our experiment, while sensitive in general to the balance of value and cost, were not *fully* resource-rational with respect to the specific parameters of their task environment.

Experiment 2

Procedure

We conduct a second behavioural experiment, as a small variation on Experiment 1. Rather than varying utility function texture, we now vary the decision noise of the target agent. Participants (total $n = 30$) were divided equally between a 'low noise' condition, where they encountered an agent with $\beta = 0.01$, and a 'high noise' condition, where $\beta = 1.0$. The game structure and mechanics were otherwise unchanged from Experiment 1. All utility functions were taken from the 'smooth' condition of Experiment 1 (lengthscale = 2.0), and absolute map costs were kept constant between all participants.

Results

The results of Experiment 2 are shown in Figure 4. Participants' selection of representations is compared to the same three models as used for Experiment 1. In this setting, the resource-rational model captures the idea that the cost-usefulness tradeoff is affected by target agent β. As an agent's decision-making gets noisier, the marginal predictive value of information about their utility function decreases—therefore the representation strategy of an agent seeking to optimise this tradeoff should be shifted towards higher aggregation as decision noise increases. Looking at Figure 4, we can see that this trend is indeed reflected in participants' behaviour, at least to some degree: for instance, the lowest aggregation level was chosen more in the 'low noise' condition, and the highest aggregation level was chosen more in the 'high noise' condition. As in Experiment 1, the resource-rational model
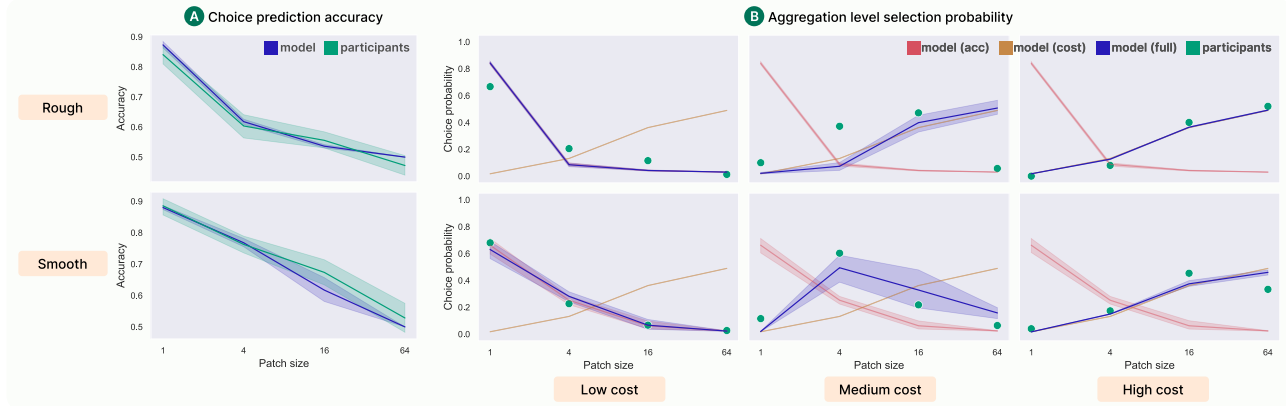
Figure 3: Results of Experiment 1. **(A)** participants' average choice prediction accuracy as a function of patch size, split by texture condition and compared to the theoretical accuracy predicted by Equation 7. **(B)** empirical patch size selection probabilities for participants from each condition, compared to those given by Boltzmann-rational models (Equation 9) based on only expected accuracy, only information cost and the full cost-adjusted return objective (Equation 8), with β = 0.25.

Figure 4: Results of Experiment 2: empirical patch size selection probabilities for participants from low and high noise conditions, compared to those given by Boltzmann-rational models (Equation 9) based on only expected accuracy, only information cost and full cost-adjusted return (Equation 8), with model β = 0.45.

gives a better fit than either of the accuracy-only or cost-only models, but again we see points of noticeable deviation (at patch size = 16 in the 'low noise' condition and patch size = 4 in the 'high noise' condition).

Discussion

In this paper, we have considered the relatively unexplored problem of how much information to represent about other agents in social cognition, through the example task of predicting an agent's choices over pairs of options. Specifically, we examined the tradeoff that an observer agent faces between information cost and predictive value in choosing how much information to represent about a target agent's utility function. We first presented some brief theoretical and computational analysis of how a simple state aggregation strategy can be used to navigate this tradeoff. We then conducted two behavioural experiments to compare people's choices of representation to a resource-rational state-aggregation model. Our findings were mixed: while for both experiments the resource-rational model fit our recorded data better than simpler decision rules based only on expected predictive accuracy *or* representation cost, participants still showed non-trivial deviations from optimality. This may be explained by the fact that our experimental setup is highly simplistic, and uses only a single *explicit* representation cost as stand-in for the real cognitive costs of information acquisition and storage. For instance, participants' behaviour may look closer to optimal under an extended model that accounts for additional constraints on e.g. attention and memory. Future experimental work should attempt to probe these nuances, and bridge the high-level computational view presented in this paper with more detailed and psychologically grounded notions of cognitive cost. An additional direction for future work is exploring strategies for obtaining resource-rational representations of agent utility functions that go beyond naive state aggregation—e.g. by representing individuals primarily in terms of their group affiliations or other social identity cues.
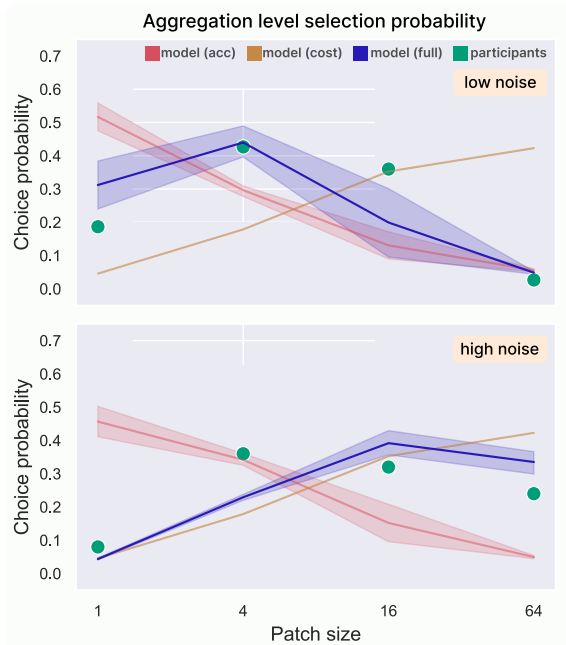
Appendix

A: Representation entropy

Here we provide some additional details on our use of representation entropy as a measure of cognitive cost. For a comprehensive introduction to entropy (and other related information-theoretic quantities) we direct the reader to Cover & Thomas (2006) or MacKay (2002). For a discrete random variable $X \in \mathcal{X}$ with probability mass function $p(x) := \Pr\{X = x\}$, the entropy of X is given by

$$H_b(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b(p(x)) \tag{10}$$

When $b = 2$ (as it typically is), the entropy has units of bits. For a continuous random variable $Y \in \mathcal{Y}$ with probability density function $p(y)$, the differential entropy of Y is given by

$$H_b(Y) = -\int_{y \in \mathcal{Y}} p(y) \log_b(p(y)) \tag{11}$$

If Y follows a multivariate Gaussian distribution with covariance Σ, then $H(Y)$ is computed as

$$H_2(Y) = \frac{1}{2} \log_2 |\Sigma| + \frac{n}{2} \left(\log_2(2\pi e) \right) \tag{12}$$

where n is the dimensionality of Y (Rasmussen & Williams, 2006). This allows us to compute the entropy of our agent utility functions u, since each is a continuous random variable distributed according to a multivariate Gaussian with known Σ. To compute the entropy of a state-aggregated utility function estimate $\hat{u}$, we can treat $\hat{u}$ as a new continuous RV with a lower-dimensional multivariate Gaussian distribution whose covariance Σ_{agg} is determined entirely by Σ and the level of state aggregation. Determining Σ_{agg} is then sufficient to compute $H(\hat{u})$.

B: Expected prediction accuracy from approximate utility functions

Let $\hat{u}$ be an arbitrary approximation to the utility function u. We want to find an expression for $\mathbb{E}_{\mathcal{S}^2} \left[\mathbb{1}(c_{pred} | \hat{u} = c_{actual}) \right]$— that is, the expected accuracy of an observer predicting the choices of a noisily rational agent over pairs of different states sampled independently from $\mathcal{S}$, given that the observer represents the target agent's utility function as $\hat{u}$. For any given pair of states (s_1, s_2) we define $\Delta u = u(s_1) - u(s_2)$ and $\Delta \hat{u} = \hat{u}(s_1) - \hat{u}(s_2)$. Since the optimal prediction strategy is to predict the higher-value state, the prediction made for a given state pair, guided by representation $\hat{u}$, depends only on $\text{sign}(\Delta \hat{u})$. For any particular pair (s_1, s_2), the sign product between u and $\hat{u}$ can take one of three values: $\text{sign}(\Delta u \Delta \hat{u}) \in \{-1, 0, 1\}$. Let p_u be the probability that an observer representing the *full* u would predict the choice correctly. We can then express the equivalent probability for $\hat{u}$ $p_{\hat{u}}$ in terms of p_u as

$$p_{\hat{u}} = \Pr\{ \text{sign}(\Delta u \Delta \hat{u}) = 1 \} p_u$$
$$+ \Pr\{ \text{sign}(\Delta u \Delta \hat{u}) = -1 \}(1 - p_u)$$
$$+ \Pr\{ \text{sign}(\Delta u \Delta \hat{u}) = 0 \} \frac{1}{2}$$

Using the fact that $\mathbb{E}_{\mathcal{S}^2} \left[\text{sign}(\Delta u \Delta \hat{u}) \right] = \Pr\left(\text{sign}(\Delta u \Delta \hat{u}) = 1 \right) - \Pr\left(\text{sign}(\Delta u \Delta \hat{u}) = -1 \right)$, we can then write the expected prediction accuracy using $\hat{u}$ as

$$\mathbb{E}_{\mathcal{S}^2} \left[\mathbb{1}(c_{pred} | \hat{u} = c_{actual}) \right] = \frac{\mathbb{E} \left[\text{sign}(\Delta u \Delta \hat{u}) \right] + 1}{2} p_u$$
$$+ \frac{1 - \mathbb{E}_{\mathcal{S}^2} \left[\text{sign}(\Delta u \Delta \hat{u}) \right]}{2} (1 - p_u) \tag{13}$$

Substituting

$$p_u = \mathbb{E}_{\mathcal{S}^2} \left[\frac{1}{\exp\left(\frac{-|\Delta u|}{\beta} \right) + 1} \right] \tag{14}$$

(from the definition of the Boltzmann-rational choice rule), and using the identity

$$(z + 1) \frac{1}{\exp(-x) + 1}$$
$$+ (1 - z) \left(1 - \frac{1}{\exp(-x) + 1} \right)$$
$$= z \tanh\left(\frac{x}{2} \right) + 1 \tag{15}$$

we obtain

$$\mathbb{E}_{\mathcal{S}^2} \left[\mathbb{1}(c_{pred} | \hat{u} = c_{actual}) \right] = \frac{1}{2} + \frac{1}{2} \mathbb{E}_{\mathcal{S}^2} \left[\text{sign}(\Delta u \Delta \hat{u}) \tanh\left(\frac{\Delta u}{2\beta} \right) \right] \tag{16}$$

References

Abel, D., Arumugam, D., Asadi, K., Jinnai, Y., Littman, M. L., & Wong, L. L. (2019, Jul.). State abstraction as compression in apprenticeship learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 3134-3142. doi: 10.1609/aaai.v33i01.33013134

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Arumugam, D., Ho, M. K., Goodman, N. D., & Van Roy, B. (2024, 04). Bayesian reinforcement learning with limited cognitive load. *Open Mind*, *8*, 395-438. doi: 10.1162/opmi_a_00132

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 1–10. doi: 10.1038/s41562-017-0064

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15-21. doi: 10.1016/j.cobeha.2021.02.015

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory (wiley series in telecommunications and signal processing)* (Second ed.). USA: Wiley-Interscience.

Craik, K. (1943). *Hypothesis on the nature of thought*.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.

Icard, T. F. (2023). *Resource rationality.* (Unpublished manuscript)

Jeffrey, R. C. (1965). *The logic of decision*. New York, NY, USA: University of Chicago Press.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*. doi: 10.1017/S0140525X1900061X

MacKay, D. J. C. (2002). *Information theory, inference & learning algorithms*. USA: Cambridge University Press.

Malle, B. F. (2008). Chapter 12 - the fundamental tools, and possibly universals, of human social cognition. In R. M. Sorrentino & S. Yamaguchi (Eds.), *Handbook of motivation and cognition across cultures* (p. 267-296). San Diego: Academic Press. doi: 10.1016/B978-0-12-373694 -9.00012-X

Neumann, J. V., & Morgenstern, O. (1953). *Theory of games and economic behavior* (Third ed.). Princeton, NJ: Princeton University Press.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, Mass.: MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second ed.). The MIT Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55 4*, 189-208.

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*, 267-279.

Model Verification and Preferred Mental Models in Syllogistic Reasoning

Sara Todorovikj (sara.todorovikj@hsw.tu-chemnitz.de)
Daniel Brand (daniel.brand@hsw.tu-chemnitz.de)
Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology
Straße der Nationen 62, 09111 Chemnitz, Germany

Abstract

A core cognitive ability of humans is the creation of and reasoning with mental models based on given information. When confronted with indeterminate information, allowing for the existence of multiple mental models, humans seem to recurrently report specific models - so-called preferred mental models. In this paper, we revisit this within the context of syllogistic reasoning, which involves statements about quantified assertions. We present an experiment designed to investigate the verification process of preferred mental models. Our analysis centers on two primary research questions: Is model verification generally straightforward for reasoners? And does a preference effect for specific models exist in syllogistic reasoning? Furthermore, employing modeling techniques, we analyze the structural complexity of mental models, based on the types of instances they consist of. We discuss our findings and their implications on the differences between reasoning with syllogisms and spatial statements.

Keywords: Mental Model Theory; Preferred Mental Models; Syllogistic Reasoning; Individual Differences.

Introduction

Consider the following reasoning example:

> All blue shapes are circles.
> All blue shapes have a diamond mark.
>
> What, if anything, follows?

This problem is a so-called syllogism. The task at hand is to determine what kind of relation, if any, exists between the two end-terms, *circles* and *diamond (mark)*, also called subject and predicate, respectively. In general, a syllogism is defined by its quantifiers (*mood*) and term order (*figure*). We take into consideration these four first-order logic quantifiers: *All* (A), *Some* (I), *Some...not* (O) and *None* (E). The figure is determined by the order of the subject, middle term and predicate of the syllogism, represented by A, B and C, respectively, in the following notation (adopted from Khemlani & Johnson-Laird, 2012):

Figure 1	Figure 2	Figure 3	Figure 4
A-B	B-A	A-B	B-A
B-C	C-B	C-B	B-C

A syllogism can be denoted using the given mood abbreviations and figure, for example the syllogism above is AA4 Conclusions are denoted in a similar fashion using the quantifier's abbreviation and the order of the end-terms (*ac* or

ca), e.g. *Eca* denotes 'No C are A'. Finally, 'No valid conclusion' is abbreviated by NVC. There exist at least twelve theories that aim to explain and model the processes behind human syllogistic reasoning (for an overview, see Khemlani & Johnson-Laird, 2012). One of the most prominent theories among them is the Mental Model Theory (MMT; e.g., Johnson-Laird, 1975, 2010). MMT postulates that given some observations, individuals create iconic representations – *mental models* – of possibilities. They create their own subjective mental representation of the information presented in a reasoning task. Considering the example above, one possible representation would be:

> circles   [blue]   [diamond]
> circles

The square brackets around an instance denote that the set of entities described by it is exhaustively represented. Another possible mental model representation is:

> circles
> circles   [blue]   diamond
> ¬circles   ¬blue   diamond

where ¬ denotes negation. Both mental representations support the conclusion "Some circles have a diamond mark" - the logically valid conclusion to this syllogism. However, in order to confirm the validity, an individual should think of all possible premise interpretations and check if they hold. The expansion of the interpretation search space can make solving such problems difficult for humans (Johnson-Laird, 2008).

Preferred Mental Models

An empirical phenomenon has been reported in the literature concerning problem descriptions allowing for multiple possible models. Specifically, some models are preferred over others – such models are called preferred mental models (PMM).

Spatial Reasoning Spatial relational reasoning problems which can evoke multiple mental models, are not all created equally (Knauff, Rauh, & Schlieder, 1995; Ragni & Knauff, 2013). This has been demonstrated through model acceptance tasks, where participants were asked to decide whether a presented spatial arrangement matches a given set of indeterminate premises. Both the patterns of acceptance responses and the reaction times clearly show that some models are preferred over others and these models adhere to some simple construction principles.

Table 1: Canonical and non-canonical instances for a syllogistic premise with terms X and Y according to mReasoner (Khemlani et al., 2015), presented in Todorovikj et al. (2023)

Quantifier	Canonical	Non-canonical
All	X Y	¬X Y ¬X ¬Y
Some	X Y X ¬Y	¬X Y ¬X ¬Y
No	¬X Y X ¬Y	¬X ¬Y
Some not	X Y X ¬Y ¬X Y	¬X ¬Y

Syllogistic reasoning Todorovikj et al. (2023) investigated the model building process in syllogistic reasoning by empirically testing what kind of models individuals create when presented with syllogistic premises. They designed an experimental domain of objects described by their shape, colour and mark. The experiment they conducted presented participants with a syllogism describing such objects and prompted them to provide a visual representation of the premises by selecting and creating objects with their desired attributes. They found that 82% of the models were correct representations of the syllogism. After analyzing the response patterns and identifying the most frequent ones, the authors reported finding preferred mental models for 46 out of 64 syllogisms. Additionally, they examined whether the observed model building behavior is in line with the model building processes of *mReasoner*[1], a LISP-based implementation of the MMT (Khemlani & Johnson-Laird, 2013). During that analysis they did not find significant results that would confirm the relevance of the initially constructed models for the final conclusion, allowing for the possibility that the built models are not necessarily the ones used when reasoning.

Canonicality of Mental Models

In mathematical and computer sciences, canonicality refers to minimal representations that avoid redundancy and ambiguity while capturing the essential properties of an expression. Within the domain of mental models in syllogistic reasoning canonicality describes the necessity of possible instances (Khemlani et al., 2015). Specifically, which entities are absolutely necessary to represent a syllogism correctly (*canonical* set of instances), and which ones do not have to be present, but do not falsify the premises and therefore could possibly be included in a model (*non-canonical* set of instances). The canonical and non-canonical instances that can be used for building a model based on the LISP implementation of mReasoner are displayed in Table 1. When building a model in mReasoner, the ε parameter is used to describe the likelihood that an instance is drawn from the full set of possible instances in contrast to only the canonical one (Khemlani et al., 2015). When fitting the model to their data, Todorovikj et al. (2023) used the proportions of non-canonical instances in the model to approximate the respective ε value.

In this article, we reinforce the first definition of canonicality when we describe syllogistic models. We define a *canonical model* as the minimal representation of a syllogism and a *non-canonical model* as the opposite extreme, i.e., a maximal representation. For example, consider the syllogism AA1:

> All squares are blue.
> All blue shapes have a star mark.

Its canonical model would only consist of entities of the following instance:

> [square]   [blue]   [star]

The non-canonical model on the other hand, would consist of all these instances (examples of negations in red):

> [square]   blue   star
> triangle   blue   star
> triangle   red   star
> triangle   red   cross

Analogously, we define an *incorrect canonical model* as the minimal incorrect representation and an *incorrect non-canonical model* as the maximal one. In the following experiment and analysis we will use these definitions of canonical and non-canonical models as lower and upper bounds of a model's complexity and heterogeneity.

Ultimately, we pose the following two research questions that we aim to answer in this paper:

[RQ1] Is the verification of models generally easy for reasoners? How fast and accurate is that process?

[RQ2] Do preference effects for accepting models in syllogistic reasoning exist? Are certain models more likely to be accepted or rejected correctly and faster than others?

The remainder of the paper is structured as follows: We first describe our experimental design, followed by an analysis of the participants' data. Afterwards, we go in-depth with respect to the structural properties of the models and outline a regression model based on them. We conclude with a discussion of our results.

Experiment

In the experiment we conducted, participants were shown a set of syllogistic premises, followed by a visual description of a model corresponding to the syllogism, which they were asked to accept or reject. Following Todorovikj et al. (2023), the syllogistic contents were object descriptions in terms of their *shape* (circle, triangle, square), *color* (red, yellow, blue) and *mark* (plus, star, diamond). We take into consideration only the 46 syllogisms for which a preferred mental model was found. For each one of them we created six tasks by deriving the preferred mental model (PMM), the canonical

---

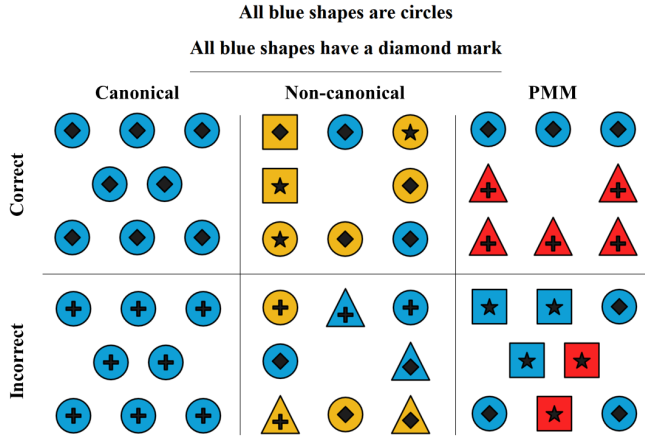[1]https://github.com/skhemlani/mReasoner

Figure 1: Illustrative example of the six models for the syllogism AA4 with contents *circle*, *blue* and *diamond*. A task in the experiment consists of two syllogistic premises describing properties of shapes and a visual representation of a specific model, as depicted here. Participants are asked to decide whether the model contradicts the premises.

model, the non-canonical model and an incorrect counterpart for each one of them, as control.

The PMMs were directly obtained from the experimental data of Todorovikj et al. (2023). For the other two models we obtained the (non-)canonical instances for both quantifiers, following Table 1, merged them based on the middle term when possible, while ensuring that they do not falsify the premises. If a merge is not possible when deriving the canonical model, then a non-canonical instance is introduced. In the case of multiple potential merges, one that minimizes negativity is chosen. That is motivated by the *principle of truth* (Johnson-Laird, 1983, 2008) which states that mental models are constructed to represent what is true according to the reasoner. We illustrate this process taking the syllogism AI2 and its canonical model as an example:

$$\text{All B are A.} \quad \text{Some C are B.}$$

The canonical instances for each premise are:

$$\text{A B} \quad \begin{matrix} \text{B C} \\ \neg\text{B C} \end{matrix}$$

The instance AB can be immediately merged with BC into ABC, which does not falsify the premises. For ¬BC, we look into the non-canonical set of the first premise, which contains A¬B and ¬A¬B, both eligible to merge with ¬BC, making A¬BC and ¬A¬BC potential instances to be added to the model. Since neither falsifies the model, we pick the one that minimizes negativity, in this case, A¬BC. Finally, the canonical model for the syllogism AI2 consists of the instances:

$$\begin{matrix} \text{A} & \text{B} & \text{C} \\ \text{A} & \neg\text{B} & \text{C} \end{matrix}$$

Regarding the incorrect canonical and non-canonical models, we first derive all possible incorrect models for each syllogism. We then pick the one with the least amount of unique

instances as the incorrect canonical model, and the one with the most as the incorrect non-canonical model. Similarly to above, if there are more than one possible choices, the one that minimizes negativity is chosen. For completeness, we also derive incorrect counterparts for the PMMs, by going through all possible incorrect models for a syllogism and employ a simple distance metric that measures the amount of different instances between two models. The most similar incorrect model is then chosen as an "incorrect PMM".

Every visual representation of a model consists of eight instances, since that is the maximum number of instances, should each one of them be different. If a derived model has less than eight instances, then some of them are repeated. In that case, we repeat the instances uniformly, while minimizing negativity, so that no bias is introduced because one instance appeared more than another. For the PMMs, we obtained the observed proportions by Todorovikj et al. (2023) and scaled them to our scenario of eight instances.

We found that three syllogisms have equal PMMs and canonical models (AA3, AI4 and EI3), so they have five corresponding tasks instead of six[2]. In total, we ultimately have $43 \times 6 + 3 \times 5 = 278$ tasks for 46 syllogisms. The participants are divided in five groups based on which syllogisms they are presented with. Following Todorovikj et al. (2023), we maintain a similar experience between participants by dividing the syllogisms in five groups based on their "preferedness", i.e. the construction frequency of the PMM. The final sets were then created by selecting one syllogism from each preference group, while ensuring that two syllogisms with a same quantifier order do not appear in the same set. That leads to four sets with nine syllogisms (two with 53 and two with 54 tasks) and one set with ten syllogisms and 59 tasks. The presented contents of the tasks were randomized per syllogism, per model. The resulting data and all materials are available on GitHub[3].

Table 2: Mean individual relative response time for each model and correctness. The full set consists of all tasks in the experiment, the reduced set eliminates the four syllogisms with less than 6 unique experimental tasks (AA3, AI4, EI3 and AA2).

Model	Full Set		Reduced Set	
	Correct	Incorrect	Correct	Incorrect
PMM	1.09	0.99	1.11	1.00
Canonical	1.03	0.80	1.04	0.82
Non-Canonical	1.19	0.89	1.19	0.90

[2]Due to a coding error in the experiment regarding the syllogisms AA2 and AA3, participants that answered for AA2 were not presented with its PMM and participants that answered for AA3 were presented with the same model twice (PMM = canonical). We include AA2's remaining responses when reporting response times and modeling, but not in the statistical analysis. For AA3 we only take into consideration the first appearance of the repeated task, to avoid potential, though unlikely, learning effects.

[3]https://github.com/saratdr/iccm-2024-SyllogisticPMMs

Participants

100 participants took part in our online web-experiment on the platform Prolific[4]. For the following analysis we performed a binomial test to determine an answer correctness percentage threshold (64-65%, depending on participant group, $p = .05$). We eliminated three participants whose correctness percentage was below the threshold, and two more due to technical issues. Ultimately, we have $N = 95$ participants (age 20-63, $M = 36.63$, $SD = 10.48$; 69% male). All of them were native English speakers. After completing the experiment, they received compensation of 6.75 GBP.

Procedure

At first participants are given an introductory task, where it's explained that they will be given two statements describing properties of shapes and are instructed to assume they are true. When they have read the statements, they are shown a visual representation of a set of shapes and are instructed that they will have to decide as quickly as possible if the set is in line with the statements or is contradicting them. Afterwards, the experiment starts, and participants are always presented with only the syllogistic premises at first. The experiment is self-paced, so once they decide to proceed, the visual representation of the model is shown as well, which they then have to accept that it corresponds to the premises or reject it. An example of a task is shown in Fig. 1.

Analysis

First, we analyzed to which extent the participants' responses were correct. Given a noteworthy correctness average of 91.61% with errors spread across all tasks, the verification task itself seemed to be so easy for all participants that errors can likely be accounted inattentiveness instead of a systematic mistake. Thus, we proceed with analyzing only correct answers. Note that because of that, throughout the analysis, the terms *correct* and *incorrect* always denote the properties of the respective model and do not refer to participants' response correctness. For our analysis, we rely on the response time between presentation of the model visualization and the participants' responses.

Table 3: Comparison of response times between correct and incorrect models in the reduced set using the Mann-Whitney U test. Significant p-values are marked in bold (corrected with Bonferroni-Holm method).

Model	Med. Corr.	Med. Incorr.	U	p
PMM	0.90	1.02	207562.5	**<.001**
Can	0.96	0.71	182432.5	**<.001**
NCan	1.05	0.79	180508.5	**<.001**

Annotation. Med. - Median; Corr. - Correct; Incorr. - Incorrect; Can - Canonical; NCan - Non-canonical.

---

[4]https://www.prolific.co/

Table 4: Comparison of response times between types of correct models in the reduced set using the Mann-Whitney U test. Significant p-values are marked in bold (corrected with Bonferroni-Holm method).

Models	Med. 1	Med. 2	U	p
PMM vs. Can	1.02	0.96	250408.0	**.034**
PMM vs. NCan	1.02	1.05	250798.5	**.034**
Can vs. NCan	0.96	1.05	237565.0	**<.001**

Annotation. Med. - Median; Can - Canonical; NCan - Non-canonical.

Since inter-individual differences can be substantial for response times and not necessarily reflecting the cognitive processes (i.e., the time needed to actually click on a response button), especially in online experiments, where the setup is non uniform, we standardized the recorded times for our analysis: For each task, we calculated the ratio between the respective response time and the overall mean response time of an individual. In the subsequent analysis we work with two sets of responses - the full set of all responses and a reduced set that does not contain responses for syllogisms with less than 6 unique tasks – AA3, AI4, EI3 and AA2. The first three have an equal PMM and canonical model, so a statistical comparison between those two models is generally impossible, whereas AA2 was affected by a coding error. Table 2 shows the mean individual relative response times for each correct and incorrect model, for both sets. Note that the impact of the elimination of the above mentioned syllogisms on the average times is negligible.

Focusing on the reduced set, we first examine the difference between correct and incorrect models. We can immediately notice that the incorrect canonical and non-canonical models were dismissed faster than the respective correct ones were accepted (0.82 vs. 1.04 for canonical; 0.90 vs. 1.19 for non-canonical). In the case of PMMs, though, there is a smaller difference (1.00 vs. 1.11), however, the increasing trend is still present. We tested for statistical significance in the changes using the Mann-Whitney U test and found that all differences are significant ($p < .001$), as shown in Table 3, along with the respective median values, for reference. This indicates that individuals are able to identify incorrect models faster than correct ones. This is plausible given that, for tasks with a universal quantifier involved (which are 40 out of the 46 tasks), participants can immediately reject the model once they recognize only one instance that contradicts the premises without even checking the rest, in contrast to correct models, where the whole model needs to be checked.

Next, we look into the response time differences between the three (correct) models. As intended, the canonical models represent the lower bound with 1.04 and the non-canonical ones the upper bound, with 1.19. The average response time for the PMMs lays in the middle with 1.11. Once again,

Table 5: Spearman correlation analysis between each type of instance and the mean response time for each model. Significant p-values are marked in bold (corrected with Bonferroni-Holm method). Note that instances that do not appear in a model or have a constant amount among all syllogisms lack a correlation value.

Inst.		PMM	Can	NCan	IPMM	ICan	INcan
Unique	M	3.25	2.56	5.93	2.90	1	7.72
	ρ	.15	.15	-.08	.03	–	-.26
	p	**.003**	**.001**	.559	1	–	**<.001**
Nec.	M	3.73	8	4.17	2.68	2.19	2.58
	ρ	-.05	–	.06	-.09	.06	.17
	p	1	–	1	.328	1	**<.001**
U. Nec.	M	1.57	2.56	2.56	0.86	0.27	2.37
	ρ	.07	.15	.04	-.07	.06	.01
	p	1	**.001**	1	1	1	1
Poss.	M	4.27	0	3.83	3.99	0.59	3.32
	ρ	.05	–	-.06	.18	.11	.18
	p	1	–	1	**<.001**	.075	**<.001**
U. Poss.	M	1.68	0	3.37	1.53	0.07	3.25
	ρ	.11	–	-.13	.15	.11	.18
	p	.056	–	**.013**	**.004**	.075	**< .001**
Inc.	M	0	0	0	1.34	5.22	2.10
	ρ	–	–	–	-.13	-.11	-.31
	p	–	–	–	**.024**	.056	**<.001**
U. Inc.	M	0	0	0	0.52	0.65	2.10
	ρ	–	–	–	-.10	-.11	-.31
	p	–	–	–	.192	.056	**<.001**

Annotation. Inst. - Instance (type); Can - Canonical (model); NCan - Non-canonical (model); IPMM/Ican/INcan - Incorrect versions of the models; U. - Unique; Nec. - Necessary; Poss. - Possible; Inc. - Incorrect.

we performed a Mann-Whitney U test and determined that the difference in response times between the models is statistically significant. Individuals needed more time to verify PMMs than canonical models ($p = .034$), but less than non-canonical ones ($p = .034$). Clearly, canonical models were evaluated faster than non-canonical ones ($p < .001$). All test results, along with the medians are displayed in Table 4. Note that all p-values reported in the two tables are corrected after the Bonferroni-Holm method for multiple comparisons. These results indicate that even though the PMMs are models that were the most frequently constructed ones, the time needed to verify such a model is not exceptionally short or long, falling between the two extreme bounds. In other words, in the domain of syllogistic reasoning, we cannot conclude that the preference for creating a model is related to the verification time or even its correctness, given the accuracy of above 90% across all models reported above.

Modeling

In this section we look into the structure of the given models, specifically, the type of their instances. We differentiate between: a) instances that are *necessary* for a correct model representation of a syllogism; b) instances that are *possible* to be added, i.e. do not contradict the premises, but aren't necessary and c) *incorrect* instances. Moreover we also look into *unique* instances, disregarding repetition. We analyzed the relationship between these descriptors and the mean relative response times. Table 5 shows the correlation results. We observe how different types of instances are significantly correlated with response times of different model, e.g. the canonical models correlate with the number of unique and unique necessary instances, whereas the non-canonical ones with the amount of unique possible instances. That is coherent with the definitions of the models relying heavily on necessary and possible instances, respectively.

As a next step, we investigate whether a model of the types of instances as descriptive features can successfully represent response times for each syllogistic model. To that end, we fit 127 linear regression models with ridge regularization, using all possible combinations of features of all lengths. We selected the best one based on the lowest Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) values. Using these two metrics on a full set of models, we can determine a threshold after which the addition of parameters does not lead to a significant fit improvement, while increasing the tendency of the models to overfit, and therefore select an appropriate model. Finally, we select the linear regression model considering the amount of unique necessary ($\beta = 0.06$), unique incorrect ($\beta = -0.11$) and possible ($\beta = 0.04$) instances, with $AIC = -854.71$ and $BIC = -843.86$, achieving a mean absolute error of $MAE = 0.16$. For a more detailed comparison, we reconstructed the mean relative response times based on the times predicted by the model, as displayed in Table 6. We observe results nearly matching the true values, while preserving the increase and decrease trends among different types of models and correctness. Thus, the model highlights how the time required by individuals for model verification is heavily based on its structure. This, again, corroborates our finding that potential preferences have little effect on participants' ability to verify models. Instead, the determining factor seems to be structural complexity of the model.

Table 6: Mean predicted relative response time for each model and correctness in the full set using a linear regression model with the number of *unique necessary*, *unique incorrect* and *possible* instances as features.

Model	Correct	Incorrect
PMM	1.11	1.00
Canonical	1.02	0.83
Non-canonical	1.16	0.91

Discussion

In this article, we continue the investigation of preferred mental models in the domain of syllogistic reasoning (Todorovikj et al., 2023) by posing two research questions. Analogous to PMM evaluation in the spatial domain (Ragni & Knauff, 2013), we first examine how trivial model verification is for individuals (**RQ1**). Thereby, we designed and conducted an experiment in the same world of marked, colourful shapes. This time, participants were presented with a syllogism and a corresponding model and asked to verify whether it is in line with the premises or contradicts them. We created six tasks per syllogism by deriving their canonical and non-canonical models as lower and upper bounds, respectively, using the already determined PMMs by Todorovikj et al. (2023), and deriving incorrect versions of all three model types, as control tasks. We found that individuals accept correct models and rejected incorrect ones with an average success of 91.61%, indicating that they do, in fact, verify models with ease, regardless of their structure.

For mReasoner (Khemlani & Johnson-Laird, 2013) and the Mental Model Theory in syllogistic reasoning, these findings have two implications: First, since models seem to be easily built and verified by human reasoners, the assumption that these processes do not involve errors is confirmed by our findings. Second, however, the fact that participants don't seem to need much effort for verification and construction, also raises the question, if the model manipulation during the search for counterexamples proposed by mReasoner is plausible: After all, an alternative solution could be to repeatedly rebuild different models instead.

Furthermore, we investigated whether a preference effect exists for accepting models in syllogistic reasoning (**RQ2**) by examining the response times for each model type and correctness. We found that the canonical and non-canonical models significantly represent the lower and upper bounds, as intended, while the mean response time for PMMs is in between them. Additionally, individuals needed significantly less time for rejecting incorrect models, respective to their counterparts, following the same trend of PMMs being in the middle of canonical and non-canonical models. This does not necessarily express any sort of *preference*, but seems to largely depend on the structural components of the models. Therefore, we analyzed the behavior further by describing the models using the types of instances they contain – necessary, possible, incorrect and unique – and finding significant correlations between them and the individuals' response times. Following that trace, we fit a set of 127 regression models, capturing all feature combinations, and found that the best representation uses the amount of unique necessary, unique incorrect and possible instances in a syllogistic model. Furthermore, the model was able to replicate the patterns in the data accurately, indicating that the selected structural properties are in fact sufficient.

In the empirical analysis of PMMs in spatial reasoning, Ragni and Knauff (2013) identified that the acceptance cor-rectness of models constructed according to a preferred strategy is typically higher than for (correct) models built following a different one (92% vs. 81% and 44%). They report analog tendencies in the respective required response times as well (3.8ms vs. 4.36ms and 6.41ms). Similar findings are made by Rauh et al. (2005), who examined acceptance of conclusions following from the respective PMMs in spatial reasoning. Ultimately, we can conclude that individuals struggle with identifying and veryfing models that do not coincide with a preferred model/strategy in the spatial relational domain, but a similar conclusion can certainly not be made for syllogisms. In fact, we showed that the difference in required verification time depends on how "chaotic" a given model is and is not related to what was found to be preferred models. Logically, given a model with at least one instance contradicting the premises, the faster it's identified, the faster it will be rejected. The more frequent an instance is repeated in a model, the less time is necessary to verify all instances. Finally, a major difference between spatial reasoning tasks and syllogistic reasoning is in their typical experimental designs: During the whole duration of syllogistic reasoning tasks, both premises are usually visible, while they are only shown for a short duration (and one after the other) in many spatial reasoning experiments. It is plausible, that strategies allowing to quickly integrate new premises and without much load on working memory cause a preference for certain models to be built in spatial reasoning tasks, while the necessity is not present for typical syllogistic reasoning tasks.

So, what does this mean about preferred mental models in syllogistic reasoning? A few questions for future research and investigation arise: Why are most of the found PMMs not equal to the canonical models? It points to a tendency of individuals adding instances that are not directly observed in the premises, but also not to the extent that they reach a full fleshed-out non-canonical representation. There is a potential to interpret this as a way of communicating other possibilities exist and ensuring that this knowledge is accounted for. Though, is this a trend only among "simpler" syllogisms that by default do not require a large amount of necessary instances to represent them? Ultimately, an important point to consider is whether the reported preferred mental models are in fact the mental models individuals use to reason about a syllogism in the first place. Todorovikj et al. (2023) fit mReasoner to their data to show a lack of relevance of the mental models provided by the participants for the conclusions they provided later on. We can interpret the found preferred models as "prototypes" for a syllogistic model, however, cannot conclude that they are preferred models when reasoning, as it's done in the spatial domain.

Acknowledgements

References

Akaike, H. (1974). A new look at the statistical model iden-

tification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Johnson-Laird, P. N. (1975). Models of deduction. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 7–54). New York, US: Psychology Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press.

Johnson-Laird, P. N. (2008). *How we reason.* Oxford University Press.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. In *National academy of sciences* (Vol. 107, pp. 18243–18250).

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*(1), 4–20.

Khemlani, S., Lotstein, M., Trafton, J. G., & Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2073-2096. doi: 10.1080/17470218.2015.1007151

Knauff, M., Rauh, R., & Schlieder, C. (1995). Preferred mental models in qualitative spatial reasoning: A cognitive assessment of allen's calculus. In *Proceedings of the seventeenth annual conference of the cognitive science society* (pp. 200–205).

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588. doi: 10.1037/a0032460

Rauh, R., Hagen, C., Knauff, M., Kuss, T., Schlieder, C., & Strube, G. (2005). Preferred and Alternative Mental Models in Spatial Reasoning. *Spatial Cognition and Computation*, *5*, 239–269. doi: 10.1080/13875868.2005.9683805

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Todorovikj, S., Brand, D., & Ragni, M. (2023). Preferred mental models in syllogistic reasoning. In *Proceedings of the 21st international conference on cognitive modeling.*

Modeling the Role of Attachment in the Development of Reciprocity and Generosity

Elpida Tzafestas

Laboratory of Cognitive Science, Department of History and Philosophy of Science,
National and Kapodistrian University of Athens, University Campus, Ano Ilisia 15771, GREECE
http://scholar.uoa.gr/etzafestas/
e-mail: etzafestas@phs.uoa.gr

Abstract

We posit that early attachment to kith and kin has a marked influence on later life reasoning and especially on generosity toward others. We present a series of computational experiments showing that the final (adult) levels of generosity differ depending on such early life exposure, and this independently of the otherwise homogeneous reasoning behavior. The benchmark experiment defines a developmental progression of three stages, from attachment only to perception of cooperative or non-cooperative actions of others in a controlled social environment to finally complex environments of arbitrary participants. We use as a behavioral basis the well-known Iterated Prisoner's Dilemma game and its classic strategy Tit-For-Tat in a simulated society of individuals. We show that in the final stage of an arbitrary complex society the social scores obtained by the developing individuals are consistently higher than the reference, undeveloped (adult) individuals, and that this is due to the developed degree of nonzero generosity. We also show that individuals with a disturbed understanding of others' emotional behavior (thus of attachment) but with intact reasoning tend to be more reciprocal and less generous in the end. On the other hand, individuals with the opposite disturbance of the understanding of reasoning but with intact understanding of others' emotional behavior tend to be far more impulsive and behave as driven by attachment only. The effect of various cognitive and social parameters on the developed behaviors is also studied. Further implications of this developmental model are finally given.

Keywords: Attachment; Attraction; Development; Reciprocity; Cooperation; Iterated Prisoner's Dilemma; Tit-For-Tat; Generosity

Introduction

From our everyday experience, we know that interpersonal relations influence our behavior as well as our understanding of the world around us. Neonatal or toddler attachment are well studied by psychology (Bowlby, 1975; Mooney, 2009) and are shown to influence an individual's behavior later in life. Psychology has also been traditionally interested in the subject of 'attraction', where (dyadic) attraction indicates affect, as well as its opposite 'repulsion' and their effect on attitude, for example whether behavioral or personality similarity is a cause or an effect of attraction (starting from the pioneering work of Byrne, for example Byrne, 1969 ; Montoya & Horton, 2020 ; Wetzel & Insko, 1982). Psychology is also interested in the personality roots of attraction (for example, Montoya & Horton, 2004), in the relation of attraction with social identity (for example,

Turner et al. 1983) and in other assorted issues. We believe that all these are instances of a general attachment/attraction mechanism that may also include friendship (Hruschka & Henrich, 2016), habituation (Davies et al., 2011), interpersonal commitment (Back & Flache, 2008) and other such phenomena. All these phenomena have in common that the effect of attachment/attraction is beneficial to the social interaction and to the participating agents. Any mechanism of attachment/attraction (from here on simply 'attachment'), whether in the narrow or in the broad sense, will by definition be outside the realm of rational decision making, and it will instead constitute a reactive component capable of responding fast and at low cognitive cost to conditions of the social environment. This can be a very handy tool for human behavior, especially in hostile, harsh or stressful natural environments.

Moreover, we believe that attachment, as well as other such reactive mechanisms, is related to reasoning and rationality. One idea is that everyday reciprocal reasoning could emerge from the interaction of attachment and social imitation. We explore this idea in a computational experimental setting based on the classic cooperation problem that is modeled as an iterated prisoner's dilemma (IPD, see next). We start from the observation that newborn and very young individuals do not seem to possess or to demonstrate intricate reasoning abilities and especially reciprocation behavior, but these appear to be learned gradually. On the other hand toddlers seem to differentiate other people as "in-group" to whom they are attached (typically kin) and "out-group" toward whom they behave differently (typically strangers whom they tend to avoid).

We therefore hypothesize that young individuals imitate the behavior of the adults in their environment, and more specifically they imitate how often the latter tend to cooperate with in-group or out-group (i.e., attached or not). A subsequent stage of their development (at around 4-5 years) allows them to mind-read others and therefore differentiate between cooperative and non-cooperative actions of third parties. At that point they start imitating how often others cooperate when their in-group cooperates or defects, and accordingly when out-group cooperates or defects. We want to show that this developmental process allows reciprocity to emerge (cooperate when the other cooperates, and defect when the other defects) with added generosity – which is beneficial for social behavior and little understood. When individuals later enter a general, complex environment of others with very diversified and possibly

opposing interests and even at times malicious, their earlier developed reciprocity with generosity comes at play.

We are modeling this developmental process as a 3-stage process: during the **first stage** individuals respond only to in-group and out-group, during the **second stage** they perceive third parties' cooperative/non-cooperative behavior toward in-group or out-group and during the **third stage** the social environment in not controlled anymore (like a family or a school environment) but may encompass arbitrary individuals and behaviors. It is important that the learning/imitation rate decreases from stage to stage, but this will be discussed later. This staged process parallels and is loosely derived from Piaget's (1947) theory of child development, although our focus is on attachment and on emotional development and not on 'cognitive' development. Our stages correspond broadly to preschool age, primary school age and adolescence. Our methods purport to indicate abstract and general tendencies and directions of development, rather than to accurately model children's behavior.

Basic experimental setup

We experiment using agent-based simulation (Namatame & Chen 2016) of a population of agents that represent individuals of 2 types: adults (that have fixed behavior) and minors or developing individuals that imitate adults and use these learned values in their own behavior. General parameters are 100 adults with 110 minors, all connected in an attachment network with connectivity factor from 10 to 20 (i.e. each agent is attached to 10 to 20 others). In each round each adult interacts with a random other adult, minors imitate a number of adults (usually 2, selected with 75% probability from their in-group and 25% from their out-group) and finally each minor interacts with a random other minor. Each pairwise interaction is a cooperation game, traditionally modeled as a special two-party game, the Iterated Prisoner's Dilemma (IPD). In this game, two agents interact for a number of cycles (here 50) and in every cycle each agent may either cooperate (C) or defect (D). It is then assigned a score defined as follows.

Individual	Opponent	Score
C	C	3 (= Reward)
C	D	0 (= Sucker)
D	C	5 (= Temptation)
D	D	1 (= Punishment)

The first notable behavior for the IPD designed and studied by Axelrod (Axelrod & Hamilton, 1981; Axelrod, 1984) is the Tit For Tat behavior (TFT, in short) : *Start by cooperating, then return the opponent's previous move.* This behavior has achieved the highest scores in early tournaments and has been found to be fairly stable in ecological settings. TFT demonstrates three important properties, shared by most high scoring behaviors in IPD experiments: (i) it is good (it starts by cooperating), (ii) it is retaliating (it returns the opponent's defection), and (iii) it is generous (it forgets the immediate past if the defecting opponent cooperates again). In our experiments, unless

otherwise stated, all adults are TFT during the 'training' stages (first and second) and can be any behavior during the realistic third stage (we are mainly using a mix of TFT with ALLC and ALLD, that always cooperate or always defect, respectively). Note that we adopt the noisy version of IPD in which there is a nonzero probability that an agent's action will be switched to the opposite, i.e. from *COOPERATE* to *DEFECT* or vice versa. It has been shown that retaliating strategies such as TFT can score quite badly in the presence of noise, despite their superiority in the non-noisy domain (Kraines & Kraines, 1995). This happens because even accidental defections may lead to a persistent series of mutual defections by both players, thus breaking cooperation. This is what makes some degree of explicit generosity necessary to account for opponent's misbehaviors.

Attachment model

Our attachment mechanism relies on our everyday experience that people tend to be good and cooperative with the ones they are attached to and tend to be "regular" with the rest. This translates in our model as:

If (attached to the opponent) then play ALLC (always cooperate) with a probability P (here set arbitrarily to 75%).
In all other cases play as usually (for example, TFT)

During the first stage, all adults are TFT and minors or developing agents use two probabilities for cooperation with in-group (attached) and out-group (others), P(A) and P(N), respectively. These are initialized randomly (usually uniformly in the interval [0,1], but other options have also been studied without significance differences in results) and are updated during imitation as follows:

$P(A)_{new} = P(A)_{old} + rate * (P(A)_{model} - P(A)_{old})$,

where $P(A)_{model}$ is the perceived probability of cooperation of the adult role model within in-group, and accordingly for $P(N)_{new}$. This probability is computed by the minor agent as the proportion of times where its adult model has cooperated. The imitation rate is high in the first stage (here 50%) and drops in the next two stages to 20% and 10%.

During the second stage, minor or developing agents use four probabilities for cooperation with in-group cooperators and defectors and out-group cooperators and defectors, P(C|A), P(D|A), P(C|N) and P(D|N), respectively. These are again initialized randomly as before and are updated during imitation in a similar manner as the P(A) and P(N) probabilities of the previous case.

Benchmark experiment

According to the above, we define our 3-stage experiment as follows. During the first stage we have adult TFTs and developing agents that can only identify and respond to in-group and out-group, while during the second stage the developing agents can also identify and respond to cooperative and non-cooperative moves of third parties and thus of their role models of imitation. Finally, during the

third stage, adults become a mix of IPD strategies (TFT, ALLC, ALLD) and minor or developing agents do not change, except for the decreasing learning rate.

Table 1 summarizes the results in presence of noise. All results are averages per agent over 50 experiments. The theoretical maximum score for cooperative agents is 150 (50 rounds by reward 3 per round). Minors develop from initial random cooperative behavior to almost full in-group cooperation and over 50% out-group cooperation (95% and 57%, respectively, after stage 1), to about 90% reciprocal cooperation independently of attachment but with very high generosity for in-group compared to that for out-group (75% compared to 10%) or equivalently close to TFT behavior for out-group (stage 2), to finally a little higher in-group cooperation (90% and 79%, respectively) with prudent and generous out-group reciprocity (64% and 33%, as C response to C and D), which is beneficial since 'behavioral noise' is present in an arbitrary population mix of ALLC, ALLD and TFT (stage 3). It is also noteworthy that in stage 1 the minors converge to a behavior whose score matches that of the adult TFT agents that they imitate but they are clearly superior in subsequent stages. This is apparently due to the high degrees of generosity that they demonstrate.

Table 1: Basic results in presence of noise. Rows correspond to the developmental stages thus each column shows the progression from stage 1 to stage 3. The two first columns show average scores for adults and developing individuals. The next four columns show the probabilities of cooperation, where A = attached, N = not attached, C = opponent cooperated, D = opponent defected. All numbers are averages of 50 experiments.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
118.912	117.752	0.952		0.566	
120.838	138.23	0.89	0.747	0.895	0.11
109.477	126.455	0.898	0.79	0.637	0.325

Note that the effect of attachment is underestimated in the previous results because an agent's in-group is much smaller than its out-group, therefore the contribution of in-group to final score is smaller than that of out-group.

These results also depend crucially on the ability to learn at all stages. In principle, this setup corresponds to enough time being available at any stage. But this is not the only imaginable setup. We discuss several variations next.

Developing socially

In practice we know of many occasions when the development environment at young age or the family is disturbed, arbitrary or not protective. Some argue that the sociocultural environment has a bigger effect on emotional development than the biological cognitive development process (for example, Vygotsky 1978 and followers).

Table 2 gives the same results as table 1 in the case of ALLD adult agents in the first two stages of development. In the third stage, a behavioral mix of agents is presented as

before. When there is a small degree of learning in the third stage, even if much lower than the previous stages, the final behavior and attachment profile stabilizes to the same as with normal development (with TFT agents in the first two stages). But if exposure to the general, behaviorally mixed, environment starts at a late stage, when learning has stopped, the results are dramatic. The agents have learned to be unresponsive to their partners' behavior, and highly cooperative (75%) when attached to a partner but fully defecting when not attached –this appears like a childish, not rational, behavior. Naturally, the scores in this case are very low.

The results obtained also differ from those of table 1 if the development environment in the first two stages consists of otherwise "rational" (TFT) agents but that are not attached to others, as table 3 shows. The minors develop as significantly less generous for attached partners, almost without discrimination between cooperators and defectors (70% and 66%, respectively, instead of 90% and 79% in the regular case).

Table 2: Same as table 1, but will ALLD adult agents during the first two stages, and a behavioral mix in the third stage. Rows correspond to the developmental stages thus each column shows the progression from stage 1 to stage 3. (a) Learning rates for the three stages: 0.5, 0.2 and 0.1, respectively.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
72.274	73.447	0.749		0.0	
71.893	72.755	0.749	0.75	0.0	0.0
114.039	124.74	0.83	0.797	0.65	0.332

(b) Learning rates for the three stages: 0.5, 0.2 and 0, respectively.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
72.161	72.068	0.503	0.505	0.75	0
72.34	72.234	0.75	0.751	0	0
114.146	71.955	0.75	0.751	0	0

Table 3: Same as table 1, but without adult attachment during the first two stages. Rows correspond to the developmental stages thus each column shows the progression from stage 1 to stage 3.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
116.205	115.381	0.5		0.542	
115.802	136.363	0.5	0.5	0.891	0.109
113.645	124.189	0.7	0.662	0.655	0.339

The attachment profile that develops does not differ, however, if the development environment contains a mix of agents and not only TFTs, provided that there is "enough" attachment at all stages. If the attachment factor drops from stage to stage, that is if developing agents are less and less attached to others as their social circle changes, the minors develop as less generous with their attached partners. These are shown in table 4. Note that minors' behavior develops normally and is not disturbed by lower attachment factor in

later stages if the adult agents of reference at the stages 1 and 2 are TFT, thus rational.

Interestingly, if noise is absent during at least one of the initial developmental stages, the minors develop as much more generous with their attached partners. These are shown in table 5.

Table 4: Same as table 1, but with a behavioral mix of adult agents during all three stages. Rows correspond to the developmental stages thus each column shows the progression from stage 1 to stage 3.
(a) Regular attachment factor during all stages (up to 20 attached partners).

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
112.152	117.611	0.889		0.516	
113.398	125.284	0.804	0.742	0.652	0.336
113.029	126.283	0.858	0.792	0.652	0.335

(b) Regular attachment factor during 1st stage, lower attachment factor during stages 2 and 3 (up to 10 attached partners).

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
113.759	116.909	0.893		0.52	
111.461	125.045	0.698	0.657	0.65	0.333
109.966	124.441	0.761	0.712	0.653	0.341

Table 5: Same as table 1, but will no noise during stage 1 or stages 1 and 2. Rows correspond to the developmental stages thus each column shows the progression from stage 1 to stage 3.
(a) Noise absent during stage 1, but present during stages 2 and 3.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
150.0	150.001	1.0		1.0	
118.374	139.221	0.986	0.884	0.893	0.111
110.75	125.098	0.966	0.875	0.651	0.334

(b) Noise absent during stages 1 and 2, but present during stage 3.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
150.0	150.002	1.0		1.0	
150.0	149.999	1.0	1.0	1.0	1.0
110.827	125.318	0.977	0.961	0.653	0.337

These results show that attachment is a very important, even defining component of the emotional development of minors and of the degree of demonstrated interplay between reciprocity and generosity, as modulated by attachment, in later stages at life. The presence of adult role models that show attachment to others, even if they are not fully "rational" and reciprocating, is of utmost importance. How attachment structures change through the developmental stages and the degree of openness that allows learning are also factors that influence later development, It may be also thought that the influence of some of those factors feeds back into development, for example high learning in an unfriendly environment may hinder and slow down learning

next, and attachment to adults that are not attached to others may lead to fewer attachment relations in future. All these are supported by our results so far.

Behavioral extremes

We are also interested in the boundaries of development itself and in the behavioral disorders that may result from flaws of social perception. More specifically, we study the case where third party in-group and attachment is not correctly perceived due to lack of emotional understanding of others, as in disorders in the autistic spectrum (Baron-Cohen et al. 2009). We also study the case where cooperation/defection and thus reason is not correctly perceived due to social cues misunderstanding, as in some retardation disorders (Zigler & Hodapp, 1986). The latter case resembles the initial attachment-only stage, where there is no biological defect, but where the social environment is too complex for the developing agent to grasp what a third party regards as cooperative and what not or because this mind-reading ability has not developed yet. Tables 6 and 7 give the results in these two cases.

The agents with lack of in-group understanding cannot perceive whether third parties are attached to one another but can understand when these third parties respond to cooperation or defection at stage 2. These agents reciprocate highly and independently of in-group, while achieving the same score levels as the typical ones. They actually reciprocate as typical agents against out-group (compare row 3 of tables 1 and 6). On the other hand, the agents with lack of perception of reason cannot perceive whether third agents have responded to cooperation or defection but they can understand whether these third parties are attached to others or not. These agents are almost full cooperators for in-group and over 50% cooperators for out-group, and these independently of opponents' behavior. Because cooperation perception is crucial for systems like our own, with this defect the agents achieve lower scores than in the previous cases, but still a little higher than the reference adult population of the third stage.

Table 6: The same experiment of table 1, but for agents that cannot perceive attachment and in-group (see text).

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
120.31	111.225	0.509		0.492	
119.475	139.761	0.9	0.186	0.9	0.186
115.468	124.716	0.682	0.367	0.682	0.367

Table 7: The same experiment of table 1, but for agents that cannot perceive third-party cooperation (see text).

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
120.665	119.939	0.947		0.566	
121.11	120.782	0.877	0.879	0.573	0.573
111.68	118.248	0.893	0.892	0.548	0.548

In extreme behavioral cases as the above, not all abnormal social settings have a direct effect. When the minors do not

perceive reason, they do not develop "normally" and they end up less or more generous respectively toward in-group when attachment or noise is not present in the initial stages of development, as tables 8 and 9 show. Interestingly, and as expected, minors that do not perceive attachment are not affected by any of these social abnormalities.

Table 8: Same as table 7, but without adult attachment during the first two stages. Rows correspond to the developmental stages thus each column shows the progression from stage 1 to stage 3.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
115.649	115.292	0.489		0.539	
116.368	115.158	0.489	0.489	0.542	0.542
112.844	115.937	0.734	0.734	0.513	0.513

Table 9: Same as table 7, but without noise during the first two stages. Rows correspond to the developmental stages thus each column shows the progression from stage 1 to stage 3.

Adults	Minors	In-group (A)		Out-group (N)	
		P(C)	P(D)	P(C)	P(D)
150.0	149.988	0.999		1.0	
150.0	150.0	1.0	1.0	1.0	1.0
111.965	115.602	0.937	0.937	0.508	0.508

Conclusion

We have shown that a simple attachment mechanism coupled with social imitation may give rise developmentally to reciprocal behavior with generosity in complex environments. The resulting behavior and especially the degree of generosity depends crucially on the early stages of development where the individuals have limited reasoning and mind-reading capabilities. It can be shown that a succession of generations of typical, developing agents stabilizes the levels of generosity in the society. For individuals with defects of perception of in-group attachment or of third party cooperative behavior, undifferentiated "cold" reciprocity or impulsive attachment-like behavior emerges respectively. The reciprocating cooperative (reasoned) behavior appears more important than attachment, as can be deduced when one of the two components are missing, but attachment allows stable long-term relationships to build and cooperation to become spontaneous and impulsive, leaving all the cognitive reasoning potential to be used more productively. This way, agents may opt for interacting preferentially with partners they are attracted to rather than interacting randomly within society, and thus create their preferred social circle. These results have implications for typical and atypical development, but also for build-up of bonds, partnerships and associations later in life, for development of social identity and norms and for other similar matters.

References

Axelrod, R., and Hamilton, W.D. (1981). The evolution of cooperation. *Science*, *211*, 1390-96.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Back, I., & Flache, A. (2008). The adaptive rationality of interpersonal commitment. *Rationality and Society, 20*, 65-83.

Baron-Cohen, S., Golan, O., & Ashwin, E. (2009). Can emotion recognition be taught to children with autism spectrum conditions? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3567–3574.

Bowlby, J. (1975). *Attachment (Attachment and loss, Volume I)*. Basic Books.

Byrne, D. (1969). Attitudes and attraction, In: *Advances in Experimental Social Psychology, Volume 4*.

Davies, A. P., Watson, R.A., Mills, R., Buckley, C.L., & Noble, J. (2011). "If you can't be with the one you love, love the one you are with": How individual habituation of agent interactions improves global utility. *Artificial Life*, 17, 167-181.

Hruschka, D.J., Henrich, J. (2016). Friendship, cliquishness and the emergence of cooperation. *Journal of Theoretical Biology*, 239, 1-15.

Kraines, D., & Kraines, V. (1995). Evolution of learning among Pavlov strategies in a competitive environment with noise. *Journal of Conflict Resolution*, 39, 439-466.

Montoya, R.M., & Horton, R.S. (2004). On the importance of cognitive evaluation as a determinant of interpersonal attraction, *Journal of Personality and Social Psychology*, 86(5):696-712.

Montoya, R.M., & Horton, R.S. (2020). Understanding the attraction process, *Soc. Personal. Psychol. Compass*, 14:e12526.

Mooney, C.G. (2009). *Theories of Attachment: An Introduction to Bowlby, Ainsworth, Gerber, Brazelton, Kennell, and Klaus*. Redleaf Press.

Namatame, A., & Chen, S.-H. (2016). *Agent-Based Modeling and Network Dynamics*, Oxford University Press.

Piaget, J. (1947). *La psychologie de l'intelligence*. Paris: Armand Colin.

Turner, JC, Sachdev, I., & Hogg, M.A. (1983). Social categorization, interpersonal attraction and group formation, *British Journal of Social Psychology*, 22, 227-239.

Vygotsky, L.S. (1978). Mind in society: The development of higher psychological processes, Harvard University Press.

Wetzel, C.G., & Insko, C.A. (1982). The similarity-attraction relationship: Is there an ideal one?, *Journal of Experimental Social Psychology*, 18, 253-276.

Zigler, E., & Hodapp, R.M. (1986). *Understanding Mental Retardation*, Cambridge University Press.

Simulating Event-Related Potentials in Bilingual Sentence Comprehension: Syntactic Violations and Syntactic Transfer

Stephan Verwijmeren (stephan.verwijmeren@ru.nl)
Department of AI, Radboud University
Nijmegen, The Netherlands

Stefan L. Frank (stefan.frank@ru.nl)
Centre for Language Studies, Radboud University
Nijmegen, The Netherlands

Hartmut Fitz (hartmut.fitz@mpi.nl)
Max Planck Institute, Radboud University
Nijmegen, The Netherlands

Yung Han Khoe (yunghan.khoe@ru.nl)
Centre for Language Studies, Radboud University
Nijmegen, The Netherlands

Abstract

Event-related potentials (ERPs) are used to study how language is processed in the brain, including differences between native (L1) and second-language (L2) processing. A P600 ERP effect can be measured in proficient L2 learners in response to an L2 syntactic violation, indicating native-like processing. Cross-language similarity seems to be a factor that modulates P600 effect size. This manifests in a reduced P600 effect in response to a syntactic violation in the L2 when the syntactic feature involved is expressed differently in two languages. We investigate if this reduced P600 effect can be explained by assuming that ERPs reflect learning signals that arise from mismatches in predictive processing; and in particular that the P600 reflects the error that is back-propagated through the language system (Fitz & Chang, 2019). We use a recurrent neural network model of bilingual sentence processing to simulate the P600 (as back-propagated prediction error) and have it process three types of syntactic constructions differing in cross-language similarity. Simulated English-Spanish participants displayed a P600 when encountering constructions that are similar between the two languages, but a reduced P600 for constructions that differ between languages. This difference between the two P600 responses mirrors what has been observed in human ERP studies. Unlike human participants, simulated participants showed a small P600 response to constructions unique to the L2 (i.e., grammatical gender), presumably because of how this grammatical feature is encoded in the model. Our modelling results shed further light on the viability of error propagation as an account of ERPs, and on the effects of syntactic transfer from L1 to L2.

Keywords: Event-related potential; P600; bilingualism; cross-language similarity; syntactic transfer; recurrent neural network; sentence processing

Introduction

Event-related potentials in bilingualism

Electroencephalography is a technique for recording electrical voltage potentials produced by neural activity. Recorded potentials can be analyzed in relation to cognitive events, yielding interpretable patterns called event-related potentials (ERPs; Morgan-Short, 2014). ERP effects have for instance been observed in response to reading words in sentence-comprehension studies. More specifically, syntactic violations result in an increased positivity in the ERP waveform that starts at around 600 ms after observing an anomalous word, as compared to its correct counterpart (Osterhout & Mobley, 1995). This ERP effect is called a P600.

The P600 effect has been used to investigate if second-language (L2) learners show similar ERP effects as native (L1) speakers for morpho-syntactic processing. L2 proficiency is the most important factor determining P600 size (Antonicelli & Rastelli, 2022; Caffarra, Molinaro, Davidson, & Carreiras, 2015; McLaughlin et al., 2010; Morgan-Short, 2014) but similarities and differences between the L1 and L2 often modulate the effect of proficiency. Some ERP studies showed reduced P600 effects, or no P600 effect, for syntactic features that are instantiated differently between languages (Antonicelli & Rastelli, 2022; Liu, Dunlap, Tang, Lu, & Chen, 2017; Morgan-Short, 2014), while others found P600 effects for syntactic L2 features regardless of the (dis)similarity between L1 and L2 (Caffarra et al., 2015; McLaughlin et al., 2010; Morgan-Short, 2014). There appears to be a complex influence of L1-L2 similarity. Native-like L2 processing (i.e., showing a native-like P600 response) of syntactic features that are unique to the L2 is possible (Foucart & Frenck-Mestre, 2012; McLaughlin et al., 2010; Morgan-Short, 2014), as is native-like L2 processing of syntactic features that are expressed similarly in the L1 and L2 (Foucart & Frenck-Mestre, 2011; McLaughlin et al., 2010; Morgan-Short, 2014). But when a syntactic feature is present but expressed differently in the two languages, the P600 seems to be less sensitive to syntactic violation in the L2 (Sabourin & Stowe, 2008; Tokowicz & MacWhinney, 2005).

Tokowicz and MacWhinney (2005) presented native English speaking learners of L2 Spanish with Spanish sentences containing syntactic violations. There were three types of syntactic violations: verb-tense violation, determiner gender violation, and determiner number violation (see Table 1). A sentence with a *tense violation* contained a verb in the progressive tense without an auxiliary verb. The syntactic con-

Table 1: Constructions containing syntactic violations with Spanish example sentences and their English translation. Words indicated with an asterisk are experimentally manipulated (here shown in the violation condition). Critical words are underlined. Table adapted from Tokowicz and MacWhinney (2005).

Violated feature	Similarity	Example sentence Spanish	English translation
Tense	Similar	Su abuela *cocinando muy bien	His grandmother *cooking very well
Determiner gender	Unique	Ellos fueron a *un fiesta	They went to *a-MASC party
Determiner number	Different	*El niños están jugando	*The-SING boys are playing

struction for the progressive tense is **similar** between Spanish and English. In a sentence with a *determiner gender violation*, the gender of a noun phrase was switched to the incorrect gender, resulting in a violation at the following noun. This syntactic construction is **unique** to Spanish compared to English, since the English language does not express grammatical gender. In a sentence with a *determiner number violation*, the number of the determiner was switched to the incorrect number, resulting in a violation at the following noun. In both languages, plurality of a noun is expressed by an inflectional morpheme suffix on the noun. However, unlike English, Spanish also expresses plurality in the determiner preceding a noun, which makes the syntactic construction **different** from English. Tokowicz and MacWhinney (2005) found that the P600 effect was reduced (in fact, it was not statistically significant) for determiner number violations compared to the other two types, which suggests that aspects of L1 syntax affect L2 processing; a phenomenon known as syntactic transfer. Specifically, the fact that number is not expressed on the determiner in English would make native English speakers less sensitive to determiner number in L2 Spanish. The same does not apply to determiner gender because there is no English grammatical gender to transfer to L2 Spanish.

Computational models of P600 effects

Although ERPs are a useful in psycholinguistic research, their precise functional interpretation is still unclear (Beres, 2017; Kaan, 2007). Several computational cognitive models have been proposed to account for ERPs (Eddine, Brothers, & Kuperberg, 2022) although only few provide an interpretation of the P600 (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Fitz & Chang, 2019; Li & Futrell, 2023).

Fitz and Chang (2019) propose that P600 size corresponds to the amount of backpropagated word-prediction error in a recurrent neural network model of word-by-word sentence processing. They used Chang's (2002) Dual-path model to compute backpropagated error on sentences based on stimuli from ERP studies. The simulated P600 effects corresponded to the effect in humans across a wide range of studies, providing support for the hypothesis that ERPs reflect learning signals in the language system. This account of ERPs is known as the Error Propagation account.

The Dual-path model is a connectionist model of sentence production and syntactic development. The model has two pathways. The first pathway is the sequencing system that learns how words are ordered in a sentence and is based on the Simple Recurrent Network (Elman, 1990). The second pathway is a meaning system that learns how to map message content onto words in a target language. The model has also been extended to the bilingual case (Janciauskas & Chang, 2018; Tsoukala, Broersma, Van den Bosch, & Frank, 2021). Verwijmeren, Frank, Fitz, and Khoe (2023) used the Bilingual Dual-path model to simulate ERP responses to syntactic violations in L2 learning. These simulated ERPs depended on L2 proficiency in a manner that resembled human subjects, adding further support to the Error Propagation account.

The present study

We use the Bilingual Dual-Path model to investigate whether the Error Propagation account can explain the P600 results from Tokowicz and MacWhinney (2005). The model simulates native speakers of English (L1) who start learning Spanish (L2) from a later age. At every point in L2 learning, we run an experiment similar to that of Tokowicz and MacWhinney, presenting simulated participants with sentences containing a verb-tense violation, a determiner gender violation, or a determiner number violation, or with a control sentence without any violation.

Based on findings from human ERP studies (Foucart & Frenck-Mestre, 2011, 2012; McLaughlin et al., 2010; Morgan-Short, 2014), we expect a clear P600 effect of violations expressed similarly in L1 and L2 (i.e, verb-tense violations) and a clear P600 effect to grammaticality violations expressed uniquely in L2 (i.e., determiner gender violations). We expect a reduced P600 effect (in line with Sabourin & Stowe, 2008) or even an absent P600 effect (in line with Tokowicz & MacWhinney, 2005) to the determiner number violations compared to the other two violation types. The results from our simulations were largely in line with these expectations, although they did not clearly confirm our expectations for the determiner gender violations. We therefore conduct a second simulated experiment with simulated monolinguals to further explore this discrepancy. Differences between the monolingual and bilingual model predictions suggest the bilingual model does display syntactic transfer from L1 to L2.

Methods

In Experiment 1, we simulate native speakers of English who are learning L2 Spanish. We train instances of the Bilingual

Table 2: Example of an experimental sentence in all for conditions. The bold morphemes indicate the sentence position where the violation occurs.

Example sentence	Violation condition
el padre hacer -a-e una bañera	none (control)
el padre hacer **-ger** una bañera	Tense
los padre **hacer** -a-e una bañera	Number
la **padre** hacer -a-e una bañera	Gender

Dual-path model[1], using a similar model configuration as in Verwijmeren et al. (2023), to learn English from "infancy" and Spanish as L2 at a later stage. The model configuration in this paper differs from the configuration in Verwijmeren et al. (2023) in how to model's next-word prediction is fed back into the model, forming its input signal at the next time step. Following Fitz and Chang (2019) closely, the input of the current model is set to the single highest activation value of the sum of the output vector (i.e., the distribution over possible next words) and the target vector (representing the single target word) . This method emphasizes correct word prediction over actual word prediction.

The model's training input consisted of sentences in artificial versions of Spanish and English that were paired with messages that expressed their meaning. The model learned to express messages as sentences in the target language (Spanish or English) by repeatedly predicting the next word. When presented with a message, corresponding nodes in the model are activated. One of the two target-language nodes is activated, and tense and aspect nodes are activated in the Event-semantics layer. Nodes in the Concept layer are activated for content-words, and a plural node is activated for plurality of a content-word. Corresponding thematic role nodes in the Role layer are activated and fixed connections are formed with the nodes in the Concept layer depending on their thematic role.

After each training epoch, the model is evaluated to measure proficiency, and tested in experimental trials to measure simulated ERPs. For Experiment 2, we trained a monolingual Spanish model. The simulated monolingual participants are trained, evaluated, and tested in the same way as the simulated L2 learners, except that they received only Spanish.

Artificial languages and model training

The artificial languages had the same constructions as the lagnagues created by Verwijmeren et al. (2023). The two artificial languages together consisted of 259 lexical items: 121 nouns, 11 adjectives, 6 pronouns, 6 determiners, 12 prepositions, 87 verbs, 8 auxiliary verbs, 6 verb inflectional morphemes, 1 plural noun marker, and the period. Using the inflectional morphemes, verbs were generated in present or past tense, with simple, progressive or perfect aspect. Plural nouns were generated using the plural noun

marker. Plural determiners in Spanish were individual words, namely "los" and "las". For example, the semantic message: AGENT: LADY; ACTION-LINKING: CARVE; PATIENT: CAKE; AGENT-MODIFIER: OLD; TARGET-LANGUAGE: EN would be expressed in English by the sentence: "the old lady carves a cake". The semantic message AGENT: ORANGE, PL; ACTION-LINKING: DISAPPEAR; TARGET-LANGUAGE: ES would be expressed in Spanish by the sentence: "las naranja -s desaparecer -an-en".

We generated 10,000 unique message-sentence pairs for training and a different set of 200 message-sentence pairs for testing, for English and Spanish combined, for each of the 60 simulated L2 learners. The message-sentence pairs are approximately equally divided over the two languages, with the percentage of English sentences being sampled from a uniform distribution between 48% and 52% and the rest in Spanish. Sentence constructions were distributed uniformly in the training input. Following Fitz and Chang (2019), we excluded the message from 70% of the training items. Each model instance iterated five times over its monolingual English training set first, before iterating for 45 epochs over its bilingual training set. The training set's order was randomized at the start of each epoch. The model learned by steepest descent backpropagation, with momentum set to 0.9. The learning rate was first set to 0.1, it then decreased linearly to 0.02 over the 5 epochs of monolingual training, and it stayed at that value during bilingual training. The simulated monolinguals were trained in the same way as the simulated L2 learners, except that that all the message-sentence pairs were in Spanish.

Model evaluation

Linguistic proficiency of the model was tested using the 200-message-sentence-pairs test set after each epoch. Sentences produced by the model were compared to the target sentence. The model's L1 and L2 proficiency was evaluated with two accuracy measures. Following Tsoukala et al. (2021), syntactic accuracy was measured as the percentage of sentences for which all words had the correct part of speech. Meaning accuracy was measured as the percentage of sentences that are syntactically accurate and also correctly conveyed the target message without additions. As pre-registered[2], we only included the 40 simulated participants with the highest meaning accuracy in our analysis.

Differences between simulated participants

Weights are initialized randomly, and differed between simulated participants. The percentage of English versus Spanish (training and testing) sentences varied between simulated participants, ranging from 48/52 to 52/48. The distribution of constructions is the same for all simulated participants. Training, testing and experimental trial sentences in the same language with the same constructions can differ between simulated participants in two ways: singular nouns that are direct

---

[1] The model code and script for the GAMMs be accessed here: https://osf.io/nbxu6/

[2] The pre-registration can be found here: https://aspredicted .org/HSR_NKN

objects can differ in definiteness of the article, and sentences can differ in content-words resulting in different meaning of sentences. Consequently, a different content-word can result in a different grammatical gender of a noun phrase.

Experimental trials

We generated 30 Spanish control sentences to obtain simulated ERPs on. For each of the control sentences we constructed a version for every violation type (see Table 2). The control sentence was a syntactically correct, active transitive sentence. There were three violation types: (1) Tense violations, where the inflectional marker for singular verbs (-a-e) was changed to progressive verbs (-ger). (2) Determiner number violation, where the singular determiner was changed to a plural determiner. (3) Determiner gender violation, where the determiner's grammatical gender was changed. These three violations involve features that are similar to English, different from English, or unique to Spanish, respectively.

Measuring simulated ERPs

The simulated participants were tested on the experimental sentences after every training epoch. Following Fitz and Chang (2019), learning was turned on in the model while processing the experimental and control sentences, but connection weights were reset to the weights of the respective training epoch after each of those sentences to prevent learning effects during the experiment. Therefore, the simulated participants encountered each trial in the same state for all of the sentences.

We measured prediction error at the hidden layer (see Fitz & Chang, 2019, for details). The prediction error of output unit j is the difference between its activation y_j and the target activation t_j, or: $\delta_j = y_j - t_j$, with $y_j \in [0, 1]$ and $t_j \in \{0, 1\}$. In the same way as during training, error backpropagated through the network to generate error at deeper layers. Error for units connected to the output layer was calculated as shown in Eq. 1, where k indexes the units connected to the output layer with weight w_{kj}, and j references the units that are backpropagating error.

$$\delta_k = y_k(1 - y_k) \sum_{j=1}^{n} \delta_j w_{kj} \quad y_k \in [0, 1] \qquad (1)$$

Error was also calculated this way for other layers backpropagating error through the network. The simulated P600 sizes are the sums over $|\delta|$ of the recurrent-layer units. The error resulting from a violation was collected at the first position where the sentence becomes ungrammatical (see Table 2). These errors were compared to errors at the same position of control sentences.

Results

Experiment 1: simulated L2 learners

Figure 1 displays the proficiency of the model at the start and the end of bilingual training. The model learns both lan-
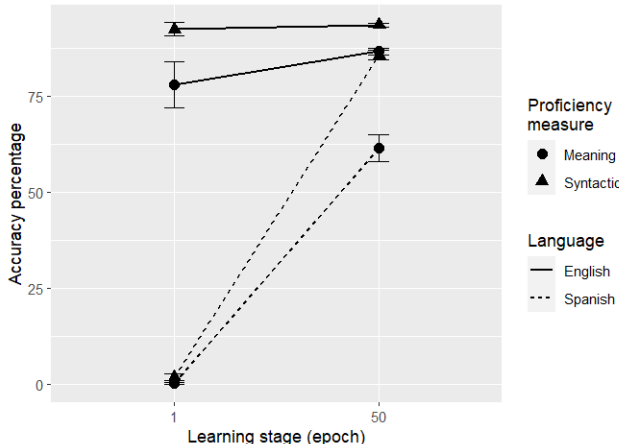

Figure 1: Mean proficiency of the bilingual model. The syntactic and meaning accuracy are displayed for the first and last epoch of bilingual training. The error bars show the 95% confidence interval.

guages to a high degree, although (unsurprisingly) it remains more proficient in L1 English than L2 Spanish.

The mean backpropagated error over L2 learning stages at the hidden layer are displayed in Figure 2. As pre-registered, we analyzed the data from our experiment with two generalized additive mixed-effects models (GAMMs; Hastie, 2017), using the bam function from the package mgcv (Wood & Wood, 2015) in R (R Core Team, 2018). Both GAMMs fit the simulated P600 effect, that is, the difference between violation and control sentences in the backpropagated error in the Bilingual Dual-path model. We fit a GAMM to determine if P600 effects differs between violation conditions Similar and Different (i.e., tense and number violations), and we fit a second GAMM to determine if P600 effects differ between conditions Unique and Different (i.e., gender and number violations).

The first GAMM[1] included the predictors of interest: DIFFERENT, LEARNING_STAGE, and their interaction. DIFFERENT indicated violation type and was dummy-coded with levels Similar and Different, coded as 0 and 1 respectively. LEARNING_STAGE is the number of L2 training epochs (standardized). We included by-participant random slopes for NOT_SIMILAR and by-participant random smooths for LEARNING_STAGE. See Table 3 (left-hand side) for a summary of the fitted GAMM. We clearly see predicted P600 effects in the Similar and Different conditions, but it is reduced in the Different compared to the Similar condition, in line with our expectations. The simulated P600 effect grows significantly over LEARNING_STAGE ($F = 33.60$, edf $= 8.61$, $p < .001$) and this growth differs between the violation types ($F = 2202.45$, edf $= 8.39$, $p < .001$).

The second GAMM[1] is the same as the first model, except for one predictor of interest, namely DIFFERENT which in this case had the levels Unique and Different, coded as 0 and 1,
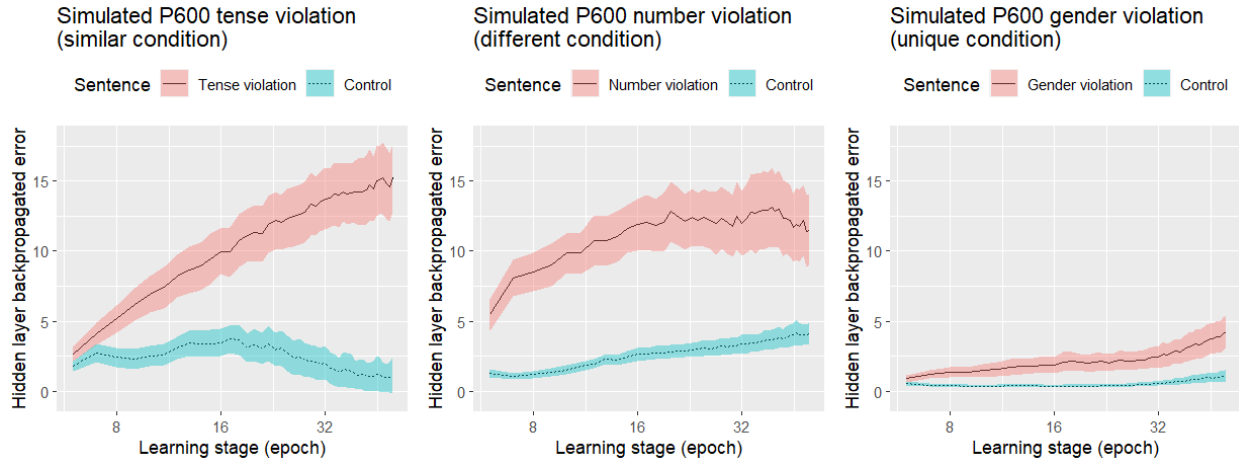
Figure 2: Mean backpropagated error (averaged over all bilingual trained model subjects) as a function of learning stage in the hidden layer, split between the three violation types. Learning stage is log-scaled. Shaded areas represent the 95% CI.

Table 3: Summary of the components in the generalized additive mixed-effects models fit on data from bilingual participants, comparing violation conditions Similar and Different (left; predictor DIFFERENT: Similar = 0, Different = 1) and the conditions Unique and Different (right; predictor DIFFERENT: Unique = 0 and Different = 1).

	Similar vs. Different				Unique vs. Different			
Predictor (coefficient)	Est.	SE	t-value	$\Pr(>\|t\|)$	Est.	SE	t-value	$\Pr(>\|t\|)$
(Intercept)	9.12	0.27	33.30	<0.001	5.26	0.267	20.83	<0.001
DIFFERENT	0.70	0.39	1.81	0.07	4.76	0.31	15.27	<0.001
Predictor (smooth)	edf	Ref.df	F-value	$\Pr(>\|t\|)$	edf	Ref.df	F-value	$\Pr(>\|t\|)$
s(LEARNING_STAGE)	8.61	8.72	33.60	<0.001	7.44	7.78	8.94	<0.001
s(LEARNING_STAGE:DIFFERENT)	8.39	8.89	2202.45	<0.001	8.79	8.98	334.19	<0.001
s(LEARNING_STAGE, participant)	295.03	359.00	48.34	<0.001	307.02	359.00	2748.53	0.05
s(DIFFERENT, participant)	77.83	78.00	447.96	<0.001	68.57	78.00	283.45	<0.001

to determine if models respond differently between violation conditions Unique (i.e., gender violation) and Different (i.e., number violation). See Table 3 (right-hand side) for a summary of the fitted GAMM. We see a weak simulated P600 effect in the Unique condition, which is smaller than the P600 effect in the Different condition. This is not in line with our expectations. The simulated P600 grows significantly over LEARNING_STAGE ($F = 8.94$, edf = 7.44, $p < .001$) and this growth differs between the violation types ($F = 334.19$, edf = 8.79, $p < .001$).

Experiment 2: simulated monolinguals

Mean Spanish meaning accuracy and mean Spanish syntactic accuracy were 99.98% and 99.99%, respectively, at the end of training.

The mean backpropagated error over learning stages at the hidden layer are displayed in Figure 3.

Similar to our pre-registered analysis, we analyzed the data from our experiment with two GAMMs, to determine if participants respond differently between conditions Similar and Different, and between Unique and Different. Both

GAMMs fit the simulated P600 effect from the Bilingual Dual-path model, here trained only on Spanish input. For the GAMM comparing Similar and Different violations, there is a larger simulated P600 effect for the Different condition compared to the Similar condition. This P600 effect significantly grows over LEARNING_STAGE ($F = 1141.37$, edf = 8.61, $p < .001$) and this growth differs between the violation types ($F = 488.73$, edf = 8.39, $p < .001$). For the GAMM comparing Unique and Different violations, there is a larger simulated P600 effect in the Different condition compared to the Unique condition. In fact, the simulated P600 effect in the Unique condition is very small. The simulated P600 effect over LEARNING_STAGE ($F = 301.10$, edf = 7.44, $p < .001$) and this growth differs between the violation types ($F = 1864.80$, edf = 8.79, $p < .001$).

Discussion

In the present work, we investigated whether syntactic (dis)similarities between L1 and L2 affect simulated L2 learners in the same way as human L2 learners. We simulated English-Spanish bilinguals and, throughout L2 learning, ex-
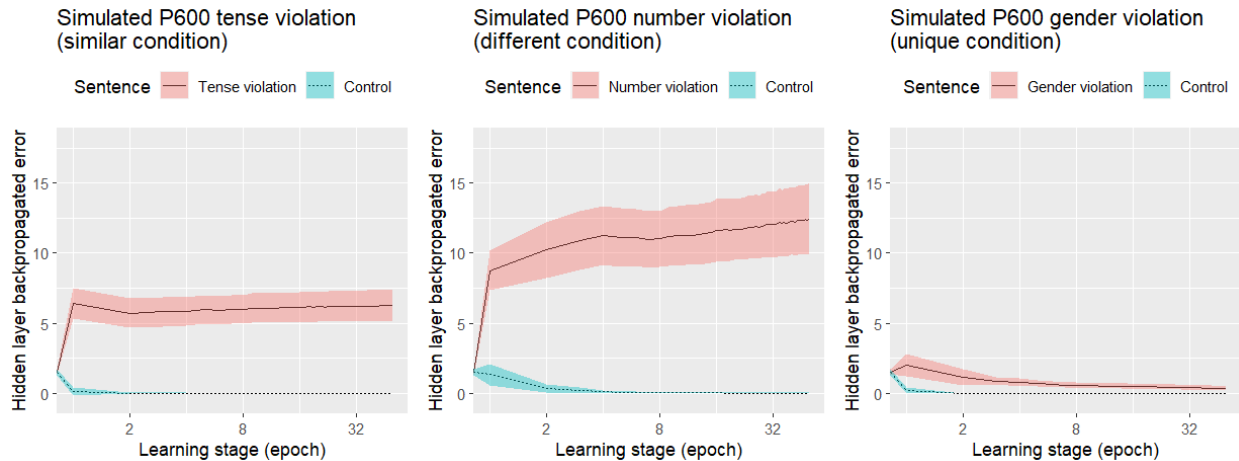
Figure 3: Mean backpropagated error (averaged over all monolingual trained model subjects) as a function of learning stage in the hidden layer, split between the three violation types. Learning stage is log-scaled. Shaded areas represent the 95% CI computed over items.

posed them to three types of syntactic L2 violations that differ in their relation to the L1. We recorded simulated P600s in response to these syntactically anomalous sentences by calculating propagated prediction error at the hidden layer, following the Error Propagation account in Fitz and Chang (2019). On this account, ERPs are summary signals of brain activity that index the propagation of prediction error during comprehension whose functional role is to support learning.

The results of our bilingual simulations are only partially in alignment with our expectations. As expected, our results reveal stronger P600 effects when syntactically anomalous sentences in the L2 contain a tense violation (similar between English and Spanish) compared to a number violation (different between English and Spanish). However, the simulated P600 effect when the L2 sentences contain a gender violation (unique to Spanish) was very weak, especially compared other two types of syntactic violations, in contrast with our expectations.

We did run our model on a simulated L1 control group and found that it predicts a *larger* P600 effect in the number violation condition compared to the tense violation condition. This is the opposite from what was found for the bilingual model's L2 and therefore support the idea that properties from the L1 affect processing in the L2 (i.e., syntactic transfer) in our model, as also appears to happen in humans (De Garavito & White, 2002; Ionin, Zubizarreta, & Philippov, 2009; Montrul, 2010; White, Valenzuela, Kozlowska-Macgregor, & Leung, 2004).

Moreover, the monolingual model showed an even smaller P600 effect in the gender violation compared to the bilingual model; an effect that reduced over L1 training whereas it increased over L2 training. Thus, it appears there is also syntactic transfer from L1 to L2 going on in the processing of gender violations.

It is not entirely clear why backpropagated error is low in response to a gender violation but not in response to a number violation. A possible explanation is the implementation of syntactic features in the model. The messages that accompany sentences during training encode tense as well as plurality of nouns, but not gender. Grammatical gender is present and expressed in our artificial language of Spanish, but there is no representation of gender in the concept layer of the model. Specifically, there is no gender node in the concept layer preceding the hidden layer, to backpropagate error to. Furthermore, verb conjugation indicating tense, as well as plurality of nouns, are expressed by morphemes that follow verbs or nouns, respectively. The model treats these morphemes as words. We have no such morphemes for gender, only separate gendered determiners for Spanish.

Conclusion

The error propagation account explained key findings from a considerable number of monolingual ERP studies (Fitz & Chang, 2019). Previous work on simulating bilingual ERPs and how they change over development (Verwijmeren et al., 2023) added further support to his account. In our present work, the reduced P600 for the number compared to tense violation supports a theory of syntactic transfer affecting ERP effects in L2 learners. The model in its present state, however, was unable to produce a strong P600 in response to a grammatical gender violation, in contrast with human participants (Antonicelli & Rastelli, 2022; Caffarra et al., 2015; McLaughlin et al., 2010; Foucart & Frenck-Mestre, 2011; Frenck-Mestre, Foucart, Carrasco-Ortiz, & Herschensohn, 2009; Morgan-Short, 2014; Tokowicz & MacWhinney, 2005). Further work is needed to determine if the Error Propagation account, as implemented in the Bilingual Dual-path model, simulates a strong P600 effect in response to a grammatical gender violation when gender is implemented in the message in the same way as plurality and tense.

References

Antonicelli, G., & Rastelli, S. (2022). Event-related potentials in the study of L2 sentence processing: A scoping review of the decade 2010-2020. *Language Acquisition*, 1–38.

Beres, A. M. (2017). Time is of the essence: A review of electroencephalography (EEG) and event-related brain potentials (ERPs) in language research. *Applied Psychophysiology and Biofeedback*, *42*, 247–255.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, *41*, 1318–1352.

Caffarra, S., Molinaro, N., Davidson, D., & Carreiras, M. (2015). Second language syntactic processing revealed through event-related potentials: An empirical review. *Neuroscience & Biobehavioral Reviews*, *51*, 31–47.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*(5), 609–651.

De Garavito, J. B., & White, L. (2002). The second language acquisition of Spanish DPs: The status of grammatical features. In *The acquisition of Spanish morphosyntax: The L1/L2 connection* (pp. 153–178). Springer.

Eddine, S. N., Brothers, T., & Kuperberg, G. R. (2022). The N400 in silico: A review of computational models. *The Psychology of Learning and Motivation*, 123.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, *111*, 15–52.

Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition*, *14*(3), 379–399.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language*, *66*(1), 226–248.

Frenck-Mestre, C., Foucart, A., Carrasco-Ortiz, H., & Herschensohn, J. (2009). Processing of grammatical gender in French as a first and second language: Evidence from ERPs. *EuroSLA Yearbook*, *9*(1), 76–106.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249–307). Routledge.

Ionin, T., Zubizarreta, M. L., & Philippov, V. (2009). Acquisition of article semantics by child and adult L2-English learners. *Bilingualism: Language and Cognition*, *12*(3), 337–361.

Janciauskas, M., & Chang, F. (2018). Input and age-dependent variation in second language learning: A connectionist account. *Cognitive Science*, *42*, 519–554.

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, *1*(6), 571–591.

Li, J., & Futrell, R. (2023). A decomposition of surprisal tracks the N400 and P600 brain potentials. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Liu, H., Dunlap, S., Tang, Y., Lu, Y., & Chen, B. (2017). The modulatory role of L1 and L2 morphosyntactic similarity during production of L2 inflected words: An ERP study. *Journal of Neurolinguistics*, *42*, 109–123.

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., & Osterhout, L. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*, *60*, 123–150.

Montrul, S. (2010). Dominant language transfer in adult second language learners and heritage speakers. *Second Language Research*, *26*(3), 293–327.

Morgan-Short, K. (2014). Electrophysiological approaches to understanding second language acquisition: A field reaching its potential. *Annual Review of Applied Linguistics*, *34*, 15–36.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, *34*(6), 739–773.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sabourin, L., & Stowe, L. A. (2008). Second language processing: When are first and second languages processed similarly? *Second Language Research*, *24*(3), 397–430.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*(2), 173–204.

Tsoukala, C., Broersma, M., Van den Bosch, A., & Frank, S. L. (2021). Simulating code-switching using a neural network model of bilingual sentence production. *Computational Brain & Behavior*, *4*(1), 87–100.

Verwijmeren, S., Frank, S. L., Fitz, H., & Khoe, Y. H. (2023). A neural network simulation of event-related potentials in response to syntactic violations in second-language learning. In *Proceedings of the 21st International Conference on Cognitive Modelling*.

White, L., Valenzuela, E., Kozlowska-Macgregor, M., & Leung, Y.-K. I. (2004). Gender and number agreement in nonnative spanish. *Applied Psycholinguistics*, *25*(1), 105–133.

Wood, S., & Wood, M. S. (2015). Package 'mgcv'. *R package version*, *1*(29), 729.

Exploring Eye-Tracking Possibilities in Algebraic Reasoning with Literal Symbols

Carolinne Das Neves Vieira (carolinne.neves@gmail.com)
Federal University of ABC, Brazil

João Ricardo Sato (joao.sato@ufabc.edu.br)
Federal University of ABC, Brazil

Abstract

Algebraic reasoning, particularly concerning literal symbols, poses significant challenges for learners and educators alike. This study investigates the potential of eye-tracking technology to enhance understanding and instructional approaches in algebraic reasoning research. Through two experimental sessions involving students aged 9-11, eye movements, and fixations were analyzed while engaging with algebraic tasks. Results reveal distinct patterns in cognitive processing, highlighting the utility of heatmaps, and eye movement videos in elucidating cognitive load and areas of difficulty. These insights inform the development of targeted instructional interventions to support learners in navigating algebraic concepts. While promising, the study acknowledges limitations in sample size and environmental control, emphasizing the need for further research. Overall, eye-tracking technology shows promise as a transformative tool in algebraic reasoning research, offering valuable insights into students' cognitive processes and informing effective pedagogical strategies tailored to the challenges posed by literal symbols.

Keywords: algebraic reasoning; literal symbols, eye-tracking technology; mathematical cognition; mathematic education.

Introduction

McNeil (2010) stated that "algebra is the foundation of higher order Mathematics and Science". Consequently, it's possible to affirm that the success or failure of algebraic reasoning may be related to the learning of an individual throughout his/her life. However, children, adolescents, and adults have faced several learning difficulties from their first contact with algebraic fundamentals. Kaput (1998) argued that algebra would be the key to Mathematics education reform. Driven by this concern, numerous studies over the years have dedicated their efforts to understanding the underlying causes of these challenges and exploring avenues for improvement. Nevertheless, few have focused on discerning the cognitive processing disparities between numerical and literal expressions.

In turn, literal symbols form the basis for understanding algebra. Pollack (2019), claims that "to learn algebra, students must develop fluency and flexibility with literal symbols". Thus, we may assume that to understand how students learn algebra we also need to understand how they process literal symbols in an algebraic context. As previously mentioned, numerous studies regarding algebra learning have been conducted and continue to be carried out, but few of them focus on understanding the differences between the comprehension of numerical symbols and literal symbols.

In Mathematics, literal symbols find application across a spectrum of contexts, serving to generalize arithmetic properties, denote unknown values in algebraic equations, represent variables in functions, and feature prominently in mathematical formulas utilized across scientific disciplines, among other uses. Given their multifaceted utility, it is unsurprising that they often engender confusion (Mcneil, 2010). This confusion can manifest in various forms, with one notably prevalent in the interpretation of algebraic expressions (Stephens, 2003).

Understanding the usage of literal symbols in mathematics can present significant challenges initially, as noted by Pollack (2019). These challenges stem from factors such as the inconsistency in the numerical magnitude of literal symbols (e.g., x could represent 6, -1, or ⅜) and the absence of a singular magnitude for them (e.g., x might denote two or four numbers or even an entire numerical set). Moreover, students' familiarity with literal symbols in literacy contexts can lead to substantial confusion when these symbols are suddenly introduced in mathematical contexts, where they are accustomed to dealing primarily with numerical representations. Thus, as Pollack (2019) observes, the cognitive processing demands associated with literal symbols are notably more intricate compared to those for Arabic numerals,

Nevertheless, while literal symbols are indispensable for comprehending algebra, the development of algebraic thinking can occur independently of their usage. As Kieran (2004) argues, although early-grade algebraic thinking may incorporate literal symbols as tools, it can also be cultivated effectively without them. This is evidenced by the ability to engage in activities such as analyzing relationships between quantities, understanding structural concepts, investigating changes, making generalizations, solving problems, modeling real-world scenarios, providing justifications, offering proofs, and making predictions.

The eye-tracking research represents a valuable tool for comprehending the diverse challenges inherent in this context. Research conducted by Bolden et al. (2015) underscores its efficacy, highlighting how "eye-tracking technology can be a useful tool in helping investigate young children's focus of attention whilst undertaking a mathematics assessment task". This methodology enables the measurement of various parameters, including the number and duration of fixations, saccades (i.e., eye movements between fixations), and pupil dilation, providing insights into cognitive processes during mathematical activities.

Andrá et al. (2015) assert that the number of fixations can serve as an indicator of how particular content is being processed. According to the author, "If an area receives a

high number of fixations, it can mean that the information is dense or complex and therefore needs to be re-examined multiple times". Consequently, differences in the number of fixations between distinct points can elucidate disparities in their processing, alongside the duration of these fixations. It is posited that stimuli of greater difficulty necessitate prolonged fixation periods.

In sum, comprehending how children reason about literal symbols in a mathematics context is essential to understanding the development of algebraic reasoning. Besides, eye-tracking technology may be a useful tool for comprehending the cognitive processing involved in children's understanding of mathematics representations. In this article, we explore the possibilities offered by this technology for studying children's cognitive processing of literal symbols and algebraic representations.

Method

Participants

In this study, we conducted two separate investigations, each involving different participants, stimuli, and days. All sessions were held in a private school in the city of São Paulo, Brazil. In the initial session, we engaged with a cohort of 9 children aged between 10 and 11 years, all in the 5th grade. Subsequently, the second session involved 9 children aged between 9 and 11 years, spanning both the 4th and 5th grades. This study was approved by the Ethics Committee of Universidade Federal do ABC and all participants and their parents signed a written consent form to participate in this study.

First Session

The primary objective of the data collected during the first session was to analyze the eye movements and quantify the number and location of fixations made by students while attempting to mentally solve four examples of five different types of tasks, as depicted in Figure 1. During the experiment, students were positioned in front of a monitor and instructed to maintain a stationary posture without moving their heads. Two calibration procedures were conducted: one before the commencement of the experiment and another at the midpoint.

Students were asked to say out loud the answer to the problem or say "I don't know" if they didn't know the answer. Their voices were recorded during the experiment to collect data on both correct and incorrect answers, allowing for a comprehensive analysis of their responses.
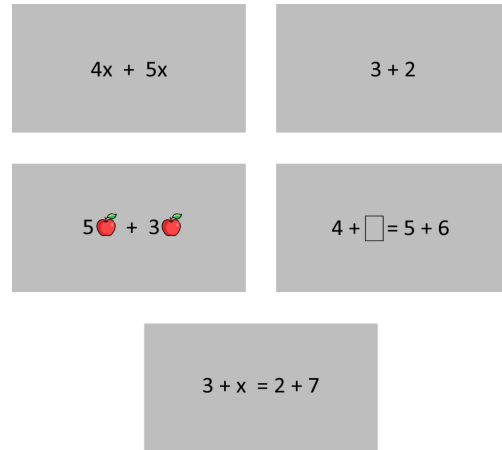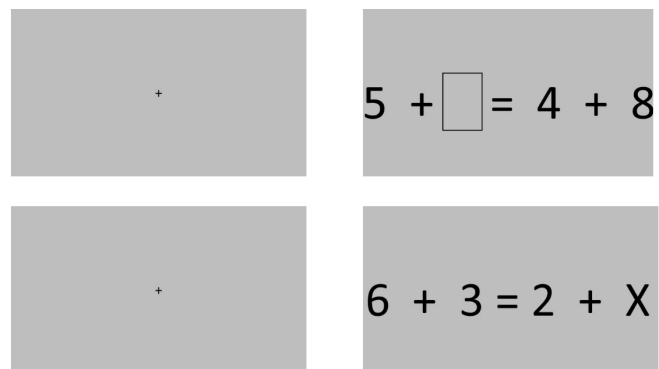

Figure 1: Types of tasks in Session 1

Second Session

In the second session, the focus of the data collecting was to analyze the number of fixations and eye movements of students while attempting to mentally solve twenty activities of each of two different types of tasks, as in Figure 2. During the experiment, students were positioned in front of a monitor and instructed to maintain a stationary posture without moving their heads. One calibration procedure was executed before commencing the experiment.

As in the first session, students were instructed to say out loud the answers for the tasks or say "I don't know" if they didn't know the answer. Their voices were recorded during the experiment to collect data on both correct and incorrect answers, allowing for a comprehensive analysis of their responses.

Figure 2: Types of tasks in Session 2

Results and Discussion

First Session

As an exploratory study, we aimed to investigate the potential applications of eye-tracking technology in algebraic reasoning research. During the first session, we observed intriguing insights using this technique. The technology provided data on the elements most fixated upon

by students, along with generating heat maps, as in Figure 5. These heat maps offer valuable insights into the areas where children direct their visual attention most time, potentially indicating areas of cognitive overload. Such insights could benefit researchers in understanding the underlying cognitive processes in children's reasoning and conceptualization, while also assisting teachers in identifying areas to prioritize during instruction.
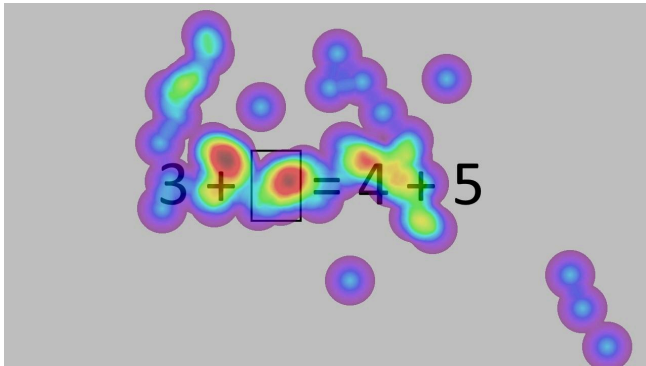

Figure 5: Heatmap indicating areas of bigger duration of fixations

Furthermore, it provides videos of the path taken by children's eyes while reasoning about the task. With these videos, we could get some insights into how this path changes depending on the type of task, as in Figure 6. To analyze these videos and comprehend the path, we established the following categories:

(1) From left to right, passing through all elements: Start from the left-hand side and move to the right, considering each element in the sequence.
(2) From right to left, passing through all elements: Begin from the right-hand side and move to the left, examining each element in order.
(3) Fixated on one point: Focus on a specific point or segment of the equation without considering the sequence.
(4) From the right-hand side to the left-hand side: Analyze the equation by moving from the right-hand side toward the left-hand side.
(5) Variable term to right-hand side to left-hand side: This sequence indicates the observation starting from the variable term, then moving right-hand side, and then to the left-hand side.
(6) Right-hand side to variable term to left-hand side: This sequence indicates the observation starting from the right-hand side, then moving to the variable term, and then to the left-hand side.
(7) Variable term to the right-hand side to variable term: This sequence indicates the observation starting from the variable term, then moving to the right-hand side and then returning to the same variable term.
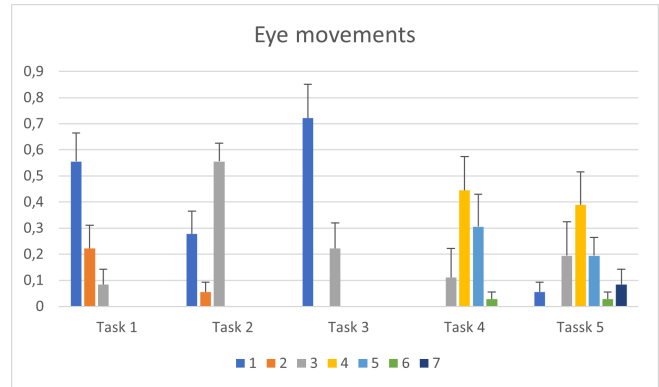

Figure 6: Changes in the direction of eye movements according to the type of task

The change is evident in the path made by students' eyes when examining expressions versus equations. We observe a distinct shift in eye movement patterns, reflecting the transition from simpler arithmetic expressions to more complex equations. This shift aligns with the previously mentioned categories, where in expressions, students predominantly follow a left-to-right path, considering each element sequentially (Category 1). In contrast, in equations, there is a noticeable tendency to analyze the equation by moving from the right-hand side toward the left-hand side (Category 4). This shift in eye movement patterns underscores the evolving cognitive processes involved in mathematical reasoning as students engage with increasingly complex mathematical concepts.

The underlying reasons for the differences in eye movement direction, as depicted in Figure 6, may stem from the cognitive strategies required for each type of task. Arithmetic expressions often necessitate a straightforward, left-to-right approach due to their simpler structure. Conversely, equations require a more integrated strategy, prompting students to scan from right to left to balance both sides and understand the relationships involved. This shift in eye movement suggests that as tasks increase in complexity, students adapt their approach, reflecting deeper cognitive processing and a more comprehensive understanding of mathematical principles.

These findings offer valuable insights into assessing children's approach to reading expressions or equations, providing a window into the cognitive processes underlying their problem-solving strategies. By analyzing eye movement patterns, researchers and educators can gain a nuanced understanding of how students engage with mathematical content. This knowledge can inform the development of targeted instructional interventions and assessment tools, ultimately enhancing mathematics education.

Second Session

During the second session, building upon the findings from the first session, our objective was to analyze videos capturing the eye movements of children as they solved

equations, along with measuring pupil diameter across two distinct types of tasks. However, technical issues arose during data collection, with some children experiencing calibration errors resulting in incomplete video data of some children.

Despite encountering technical issues, our analysis yielded intriguing heat maps. Our observations revealed a notable difference in fixation duration between blank spaces and areas with literal symbols in equations. Interestingly, while both represent the same concept, the area with the literal symbol received significantly more fixation duration compared to the area with the blank space, whereas there were almost no fixations, as in Figure 7
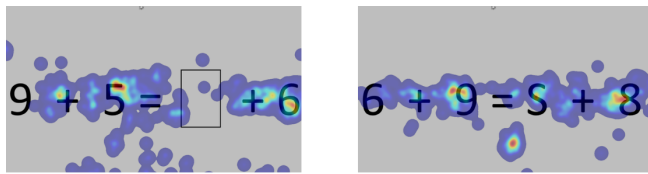

Figure 7: Heatmaps of equations

The heatmaps generated from our observations could serve as valuable tools in understanding children's cognitive processes when dealing with literal symbols in algebraic reasoning. This insight could be particularly beneficial for researchers aiming to comprehend cognitive overload during the development of algebraic reasoning, especially concerning the presence or absence of literal symbols. Additionally, educators could leverage these heatmaps to introduce algebraic concepts in a manner that minimizes cognitive shock for young learners. By utilizing the insights provided by it, teachers can better focus their instructional efforts and support students in achieving a deeper understanding of algebraic concepts.

Conclusion

Our primary goal in this article was to explore the potential of using eye-tracking technology in the research about the development of algebraic reasoning, with a focus on children's understanding of literal symbols. By providing heat maps, and videos of eye movements, this technology could serve as a valuable tool for researchers and educators seeking to improve mathematics education. Ultimately, our goal was to contribute to the advancement of teaching strategies and student learning outcomes in mathematics.

Heat maps offer a valuable means to identify elements within expressions or equations that may elicit heightened cognitive load. This insight equips researchers with a deeper understanding of children's challenges in algebraic reasoning, while also empowering teachers to target areas of difficulty more effectively. By discerning where cognitive strain occurs most frequently, educators can tailor their instructional approaches to support students' development of algebraic reasoning skills more adeptly.

Videos capturing eye movements offer a promising avenue for gaining insights into children's problem-solving approaches across various mathematical tasks, including expressions, equations, and word problems. By analyzing these videos, researchers can glean valuable insights into the strategies employed by children and the paths they traverse when tackling such tasks. Furthermore, teachers can utilize these insights to adapt their instructional strategies, providing tailored support to individual students based on their observed problem-solving approaches. By aligning teaching methods with students' cognitive processes, educators can foster a more effective learning environment conducive to improved mathematical proficiency.

In conclusion, eye-tracking technology holds immense potential as a valuable tool in mathematics education research and practice. By providing detailed perceptions of students' cognitive processes, such as problem-solving approaches and cognitive load, eye-tracking offers researchers and educators a deeper understanding of how students engage with mathematical tasks. This technology opens up a range of possibilities, from identifying areas of difficulty and cognitive overload to informing the development of targeted instructional interventions. Ultimately, the integration of eye-tracking technology into mathematics education research and teaching practices has the potential to enhance student learning outcomes and contribute to the advancement of effective pedagogical strategies in mathematics education

Limitations

We acknowledge the limitations imposed by the small sample size of our study, which precludes generalization of our findings. Further research is warranted to validate the results observed in this investigation. Additionally, the absence of environmental control during the study may have influenced certain outcomes. Conducting future studies in controlled environments, such as laboratory settings, where students can be isolated from potential distractions such as noise and other individuals, could enhance the validity of our findings and address even more valuable findings regarding this topic.

Acknowledgments

References

Andrá, C., Lindström, P., Arzarello, F., Holmqvist, K., Robutti, O., & Sabena, C. (2015). Reading mathematics representations: An eye-tracking study. International Journal of Science and Mathematics Education, (pp. 237-259).

Bolden, D., Barmby, P., Raine, S., & Gardner, M. (2015). How young children view mathematical representations:

A study using eye-tracking technology. Educational Research, (pp. 59–79).

Kieran, C. (2004). Algebraic thinking in the early grades: What is it. The mathematics educator, (pp. 139-151).

McNeil, N. M., Weinberg, A., Hattikudur, S., Stephens, A. C., Asquith, P., Knuth, E. J., & Alibali, M. W. (2010). A is for apple: Mnemonic symbols hinder the interpretation of algebraic expressions. Journal of Educational Psychology, (pp. 625–634).

Pollack, C. (2019). Same-different judgments with alphabetic characters: The case of literal symbol processing. Journal of Numerical Cognition, (pp. 241–259).

Stephens, A. C. (2003). Another Look at Word Problems. The Mathematics Teacher, (pp. 63–66).

A Neuro-Symbolic Implementation of Mouse Reward Timing Learning

Laura Sainz Villalba (laura.sainz@ini.ethz.ch)
Institute of Neuroinformatics, Winterthurerstrasse 190
Zurich, 8057 Switzerland

P. Michael Furlong (michael.furlong@uwaterloo.ca)
Centre for Theoretical Neuroscience, Systems Design Engineering
University of Waterloo
Waterloo, ON, Canada

Abstract

Animals and humans in reinforcement learning tasks are able to learn the timing of reward delivery, even when that timing is delayed and variable, suggesting a sophisticated ability to learn the distribution of reward timings. In this work, we present two algorithms simulating the switching interval variance (SIV) task as described in Li *et al.* that showed mice were able to adapt their behaviour to the change of standard deviation of the reward time delays. Both algorithms implement the wait vs stay decision by thresholding the log evidence that a forthcoming reward is likely, without assuming the specific form of the reward timing distribution. One algorithm is implemented algebraically, and the other using Spatial Semantic Pointers, a tool from Vector Symbolic Algebras for representing continuous values that have ties to hippocampal grid cells. We show that our models capture characteristic behaviour of mice on the SIV task.

Keywords: reward timing; reward learning; vector symbolic algebra; neurosymbolic programming

Introduction

Timing – adjusting behavior depending on temporal regularities of the environment – is critical for a wide range of natural behaviors, like reward timing occurrences in foraging. Previous research in conditioning and operant paradigms have demonstrated that animals learn the time delay until a reward. When reproducing specific time intervals, rodents exhibit a variance that scales with the mean time targeted (Lejeune & Wearden, 2006). This phenomenon is known as *scalar timing* or *time scale invariance*, and has driven the majority of models to explain the underlying processes for timing (Machado et al., 2009). However, in contrast to the reliable timing typical of most operant conditioning paradigms, timing of natural events can be vastly unpredictable. In general, reward timings, considered as random variables, may follow very different distributions in which the mean may not be enough to capture the variability to behave successfully. In this regard, there is mounting evidence that animals and humans are able to track measures of uncertainty to generate optimal decisions (Preuschoff et al., 2008; MacLean et al., 2012). Specifically, Li and Dudman (2013), showed mice were able to adapt their behaviour to the change of standard deviation of the reward time delays. In that paper, Li *et al.* proposed a model to reproduce this effect in which the reward timing distribution was represented by a Gaussian distribution, with recurrent estimation of parameters from empirical estimates on previous trials. However, they did not explicate a mechanism by which this could be neurally implemented.

In this paper we propose an alternative model of the waiting time in simulated mice on the same task. We formulate the choice to stay and wait for a reward as a go/no-go decision, without explicit knowledge of a distribution function. We provide two implementations: an algebraic model and an model using Spatial Semantic Pointers (SSPs) – high-dimensional representations of real-valued data that belongs to the Vector Symbolic Algebras (VSAs)[1] family of cognitive modelling tools. Building on prior work interpreting VSA statements as probabilistic statements (Furlong & Eliasmith, 2022, 2023), we translate the probabilistic algorithm implemented in our purely algebraic model to a VSA framework.

The rest of the document is laid out as follows: First, we cover the background on experimental data on timing, modelling reward timing inference and VSAs, specifically the encoding of real-valued data in VSAs (section *Background*); second, we describe the experimental set up and the algorithms used in this paper (section *Method*); third, present the results of the experiment (section *Results*); and finally we discuss the results comparing with existing models and, conclude (section *Discussion*).

Background

Models of Timing in Biological Agents

Different theories of how reward timing occurs in animals and humans have been constructed, for review see Machado et al. (2009). They share in common three components: a representation of physical time, a memory that stores information about when rewards arrive, and a mechanism to generate predictions (Yi, 2007). The focus of the majority of these models has been to explain the effect of time scale invariance, which applies to paradigms in where timing relevance can be described through the mean. In this case, the resulting variance is more related to internal noise of the animal's prediction than to external variability on the timings. Most widely known models in this regard, are scalar expectancy theory (SET) (Gibbon, 1977), Learning-to-time (LeT) (Machado, 1997) or more recently, Interval timing through neural integration (Simen, Balci, deSouza, Cohen, & Holmes, 2011).

In this work, we present a modeling approach to explain adaptive behaviour in response to changes in the variance of reward timings, and not only to the mean, which in the SIV

---

[1] Also known as Vector Symbolic Architectures

task is kept fixed (Li & Dudman, 2013). In this paradigm, the behavioural variance becomes a key variable in the optimal performance of the animal and then, it must be part of the representation itself, to predict reward timings. This variability might be represented in existing models of timing, but it has not been acted upon to the best of our knowledge.

Vector Symbolic Algebras and Probability

Vector Symbolic Algebras (VSAs) are a family of algebras over high dimensional vectors, developed in cognitive science to unify symbolic reasoning and neural networks. One VSA, the Holographic Reduced Representations VSA (Plate, 2003), have recently been reinterpreted probabilistically (Furlong & Eliasmith, 2023). The upshot of that work is that probabilistic statements can be translated directly into VSA statements, and through the Neural Engineering Framework (Eliasmith & Anderson, 2003), into populations of spiking neurons. At the heart of probabilistic VSA modelling is the recognition that dot product similarity between the high dimensional vectors can be converted into probability values. Following from that, probability distributions, $p(X = x), x \in X$, can be embedded in this vector space:

$$\mu_p = \int_X p(X = x)\phi(x)dx. \tag{1}$$

Probabilities can then be evaluated by taking the dot product between an encoded data point, $\phi(x)$, and the mean vector, μ_p, $P(X = x) \approx \phi(x) \cdot \mu_p$. Continuous values, like time, t, are represented using *fractional binding* (Plate, 1992), a method from VSAs for encoding d-dimensional representations of real-valued data. We refer to fractionally bound objects as Spatial Semantic Pointers (SSPs) (Komer, Stewart, Voelker, & Eliasmith, 2019; Dumont & Eliasmith, 2020). We construct SSPs as follows:

$$\phi(t/\lambda) = \mathcal{F}^{-1}\left\{e^{iAt/\lambda}\right\} \tag{2}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform, and $A \in \mathbb{R}^{d \times 12}$ is the phase matrix, and λ is a length scale hyperparameter. A is enforced to have conjugate symmetry, and the free elements of A are uniformly sampled from the range $[-\pi/2, \pi/2]$. However, because the dot product between SSPs can be negative, we have to map the similarity into the space of probabilities using a method from Glad et al. (2003):

$$P(X = x) = \max\left\{0, \phi(x) \cdot \mu_p - b\right\}, \tag{3}$$

where the bias, b, ensures Eq. 3 integrates to 1.

Method

We model the behaviour of a mouse when performing a switching interval variance (SIV) operant conditioning task as described by Li and Dudman (2013). For each trial, the mouse subject can choose to press one of two levers, and

---

²For m-dimensional data, $A \in \mathbb{R}^{d \times m}$, hence referring to A as a phase *matrix*.

then approach a vestibule to wait for a water reward that will be delivered after a time delay (t_{delay}). Only one of the two levers is baited and would give reward in 85% of trials. The side of the baited lever is randomly selected for each block of trials (180–200 trials/block) with block switches being uncued. The resulting design and behaviour gives rise to four trial types: a correct choice of the baited lever followed by water delivery ("correct"), a correct choice of baited lever with no water delivered ("probe"), a correct choice of baited lever with no water delivered because the mouse left earlier than the time delay for that trial ("early") and, an incorrect choice of the unbaited lever ("error"). The time elapsed from the press lever until the mouse leaves the vestibule, corresponds to the waiting time t_{wait}. If the mouse chooses to stop waiting, leaving the vestibule returning to press a lever, a new trial will be initiated. For each block, the time elapsed from the lever press to the delivery of reward, *i.e.*, t_{delay} follows a Gaussian distribution with mean $\mu = 3000$msec and one of three possible standard deviations selected at random: 50msec, 750msec, or 2000msec.

The following assumptions hold for the algebraic and SSP model implementations we present:

- The task (block and trial structure) has been already learned, assuming a state-machine structure in the elicited behaviour of the mouse subject. This includes the notion of the baited lever per block, its probability of reward, and the notion of the reinforced status of the trial. Hence, the baited status of the pressed lever is known and at each given block the learning affects only to the reward statistics.

- An ε-greedy reinforcement learning algorithm (Sutton & Barto, 2018, Ch 2) models the trade-off between exploration and exploitation. Therefore, a ε small fraction of trials corresponds to exploration of the non-baited lever.

- The distribution of t_{delay} is learned across trials adapting to the switch in distribution after an uncued transition to a new block.

We model waiting time as an evidence integration task with only one action: stop waiting for a reward at the vestibule. Our simulated subject will stop waiting when one of two conditions are met: either the reward is delivered, or the integrated evidence for no forthcoming reward meets a predetermined threshold. We integrate the probability of a waiting time, $p(t_{\text{delay}} = t_{\text{waiting}})$, into a cumulative distribution function (CDF), which we use to compute a decision function,

$$D(t_{\text{wait}}) = \log\left(\frac{P(T_{\text{delay}} \leq t_{\text{wait}})}{1 - P(T_{\text{delay}} \leq t_{\text{wait}})}\right), \tag{4}$$

that computes the log ratio between the probability that a reward should have been delivered ($P(t_{\text{delay}} \leq t_{\text{wait}})$, and the probability that the reward will be delivered in the future, $P(t_{\text{delay}} > t_{\text{wait}}) = 1 - P(t_{\text{delay}} \leq t_{\text{wait}})$, i.e. the survival function. When $D(t_{\text{wait}})$ is larger than the decision threshold, ν,

the mouse has reached enough certainty to consider that the current trial is actually a probe trial and not a rewarded one.

This model depends on a CDF for the reward delay time. In the algebraic model we base this off of a normal distribution, parameterized by μ_{delay} and σ_{delay}. In the case of the SSP model, we approximate a running average of the distribution μ_P, through a low-pass filter (LPF) on observed waiting times, as described in Section *VSA Implementation*, below. Like Li and Dudman (2013) we estimate these parameters from the last $n = 10$ trials. For probe trials we use the time the agent spent waiting, t_{wait}, as a proxy for the observed reward time.

Algebraic implementation

In the algebraic algorithm, during the waiting time and at each time step, the CDF for current waiting time is computed. The CDF value represents the probability that the reward should have been delivered up to the current time, under the current representation for the reward delay distribution. This waiting loop breaks, as explained previously, when $D(t_{\text{wait}}) \geq \nu$ or when reward is delivered.

The reward delay is modelled with a normal distribution, hence, only the sufficient statistical parameters are updated and maintained. The mean, μ, and the standard deviation, σ are estimated using the empirical mean and standard deviation of the stored waiting time samples. The update on each follows a smoothed correction weighing a delta error.

$$\Delta\hat{\mu} = \bar{\mu} - \hat{\mu}$$
$$\hat{\mu} = \hat{\mu} + w * \Delta\hat{\mu}, \tag{5}$$

$$\Delta\hat{\sigma} = \bar{\sigma} - \hat{\sigma}$$
$$\hat{\sigma} = \hat{\sigma} + w * \Delta\hat{\sigma}, \tag{6}$$

The weight depends on the reinforced status of current trial and it represents a notion of the certainty that the waiting time t_{wait} observed, could be used as a proxy for the reward delay distribution t_{delay}. If the waiting time corresponds to a trial in which reward was delivered, certainty that it is a good proxy for t_{delay} is total i.e. one. However, for the rest of non rewarded waiting times, only the ones corresponding to early trials (and not the probe trials) would represent the real statistics of the t_{delay} distribution (despite having a shifted value). Hence, notion of certainty applied in this case corresponds to the proportion that these early trials represent of all observations corresponding to non rewarded trials. When the trial is rewarded, the weighting of t_{wait} is $w_{\text{rewarded}} = 1$. When the trial is not rewarded we set:

$$w_{\text{NOTrewarded}} = \frac{P(t_{\text{delay}} > t_{\text{wait}}, r = 1 | a_{baited})}{P(t_{\text{delay}} > t_{\text{wait}} | a_{baited})},$$
$$= \frac{p_{\text{reward}} * sv(t_{\text{wait}})}{p_{\text{reward}} * sv(t_{\text{wait}}) + 1 - p_{\text{reward}}},$$

where $sv(t_{\text{wait}}) = 1 - P(t_{\text{delay}} > t_{\text{wait}})$ corresponds to the survival function and p_{reward} is $P(r = 1 | a_{baited})$. If $t_{\text{delay}} > t_{\text{wait}}$ then reward was never delivered for that trial, despite being a

rewarded trial $r = 1$. The denominator for $w_{\text{NOTrewarded}}$ represents all observed trials where the baited lever was pressed but no reward was delivered $P(t_{\text{delay}} \geq t_{\text{wait}} | a_{baited})$ under the current model for the distribution of t_{delay}. This corresponds to the sum of early trials and probe trials.

$$P(t_{\text{delay}} \geq t_{\text{wait}} | a_{baited}) = P(early) + P(probe)$$
$$P(early) = sv(t_{\text{wait}}) \cdot p_{\text{reward}}$$
$$P(probe) = 1 \cdot P(r = 0 | a_{baited}) = 1 - p_{\text{reward}}$$

$$w_{\text{NOTrewarded}} = \frac{P(early)}{P(early) + P(probe)} \tag{7}$$

This differential weighing helps to account for the fact that the sampling of the real distribution of reward delays $t_{\text{delay}} \sim \mathcal{N}(\mu_{delay}, \sigma_{delay}^2)$ is truncated, with the upper tail of the distribution being unobserved.

In this algorithm, the memory comprises of two queues of 10 trials in size (following the model in Li and Dudman (2013)) for waiting times q_{wait}, both conditioned on the subject's action. Trial history is tracked with an extra counter, $count_{\text{wait}}$, that stores the number of baited lever trials. As described in Algorithm 1, after each trial, memory is updated with resulting waiting time if the baited lever was pressed for the current trial.

VSA Implementation

Algorithm 2 gives the complete SSP implementation of the algorithm. The algorithm computes the decision function, $D(t_{\text{wait}})$, and makes the decision to leave. The algorithm computes the CDF of t_{delay}, $P(t_{\text{delay}} \leq t)$ by integrating the probability of the current waiting time. This requires representing the current waiting time, $\phi(t_{\text{wait}})$, and an estimate of the distribution over reward delay times, μ_{delay} using SSPs. In this paper we arbitrarily set the dimensionality of the SSP encoding to $d = 256$ and fix $\lambda = 1$.

To compute the probability that the reward delay time, t_{delay} is equal to the current elapsed time, t_{wait}, we need an estimate of the time that has elapsed since the simulated mouse started waiting in the vestibule. To estimate elapsed time we exploit a property of binding fractional bound quantities, that for two quantities, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, then the VSA binding operation $\phi(\mathbf{x}) \circledast \phi(\mathbf{y}) = \phi(\mathbf{x} + \mathbf{y})$. Thus, the SSP-encoding of the current time, $\phi(t_{\text{wait}})$ can be updated recursively by binding it with an SSP encoding of the simulation time step, $\phi(\Delta t)$, hence $\phi(t_{\text{wait}} + \Delta t) = \phi(\Delta t) \circledast \phi(t_{\text{wait}})$. Since binding in the HRR algebra is implemented using circular convolution, we can also write this as the product between the current time representation and a circulant matrix constructed from the vector encoding the time step, denoted Circulant($\phi(\Delta t)$) (see line 5 of Algorithm 2). This integration can be implemented by a recurrent linear network (line 4 of Algorithm 2), whose state gets reset after an action is selected by the agent. To compute the CDF, we integrate the probability observed at every time step.

211

Algorithm 1 Pseudocode for Algebraic model of waiting

Require: $q_{\text{wait}}(n_q)$ ▷ Memory for waiting times
Require: $n_q = 10$ ▷ Memory size for q_{wait}
Require: $counter(n_c)$ ▷ Memory for baited lever trials
Require: $n_c = 1$ ▷ Memory size for counter
Require: $\Delta t \geq 0$ ▷ Simulation timestep
Require: ν ▷ Decision threshold
Require: t_{arrival} ▷ Time the mouse arrives at vestibule
Require: p_{reward} ▷ Long-term probability of reward
Require: $\hat{\mu}_{\text{delay}} > 0$ ▷ Initial estimate of reward delay
Require: $\hat{\sigma}_{\text{delay}} > 0$ ▷ Initial estimate of reward standard deviation

1: $t_{\text{wait}} \leftarrow t_{\text{arrival}}$
2: **while** $t_{\text{wait}} \leq t_{\text{max}}$ **do**
3:      $p_{\text{go}} \leftarrow P\left(t_{\text{delay}} \leq t_{\text{wait}} \mid \hat{\mu}_{\text{delay}}, \hat{\sigma}_{\text{delay}}\right)$
4:      $h \leftarrow \log\left(\frac{p_{\text{go}}}{1-p_{\text{go}}}\right)$
5:      **if** $h \geq \nu \vee$ reward arrives **then**
6:          $q_{\text{wait}}.push(t_{\text{wait}})$
7:          **break**
8:      **end if**
9:      $t_{\text{wait}} \leftarrow t_{\text{wait}} + \Delta t$
10: **end while**
11: **if** baited lever press **then**
12:      $counter \leftarrow counter + 1$
13:      $w \leftarrow 1$
14:      **if** no reward **then**
15:          $w \leftarrow \frac{p_{\text{reward}}*sv(t_{\text{wait}})}{p_{\text{reward}}*sv(t_{\text{wait}})+1-p_{\text{reward}}}$
16:      **end if**
17: **end if**
18: $\mu_{\text{new}} = \text{mean}(q_{\text{wait}})$
19: $\sigma_{\text{new}} = \text{std}(q_{\text{wait}})$
20: $\hat{\mu}_{\text{delay}} \leftarrow \hat{\mu}_{\text{delay}} + w * \left(\mu_{\text{new}} - \hat{mu}_{\text{delay}}\right)$
21: $\hat{\mu}_{\text{delay}} \leftarrow \hat{\sigma}_{\text{delay}} + w * \left(\sigma_{\text{new}} - \hat{sigma}_{\text{delay}}\right)$

Algorithm 2 Pseudocode for SSP waiting model.

Require: $\Delta t > 0$ ▷ Simulation timestep
Require: $\gamma \in (0,1)$ ▷ Memory Decay Rate
Require: $\phi : \mathbb{R} \rightarrow \mathbb{R}^d$ ▷ SSP Encoding
Require: ν ▷ Decision threshold
Require: $\mu_{\text{delay}} \leftarrow \mathbf{0}$

1: $\phi(t_{\text{wait}}) \leftarrow \phi(0)$
2: $h \leftarrow 0$
3: $p_{\text{go}} \leftarrow 0$
4: **while** $h \leq \nu$ **do**
5:      $\phi(t_{\text{wait}}) = \text{Circulant}(\phi(\Delta t))\phi(t_{\text{wait}})$
6:      $p_{\text{go}} \leftarrow p_{\text{go}} + \max\left\{0, \mu_{\text{delay}} \cdot \phi(t_{\text{wait}})\right\} \Delta t$
7:      $h \leftarrow \log\left(\frac{p_{\text{go}}}{1-p_{\text{go}}}\right)$
8:      **if** reward arrives **then**
9:          $\mu_{\text{delay}} = \gamma\mu_{\text{delay}} + (1-\gamma)\phi(t_{\text{wait}})$
10:          **break**
11:      **end if**
12: **end while**

Results

In Fig. 4 we show that both the algebraic and the SSP implementations reproduce the linear relationship between reward delay standard deviation and waiting time in the vestibule although with a higher slope than observed in experimental data (dotted line). Looking at the standard error of the mean (SEM) of the algebraic and SSP implementations in Fig. 4, it appears that there is not a substantial difference in the performance of the two classes of algorithm. This is despite the fact that we are using relatively different models of representing probability distributions.

In Fig. 5, the algebraic model consistently exhibits higher correlation between waiting times of probe trials and timings of previous rewarded trials, showing the explicit dependency on the memory formulated. However, the SSP model shows low correlation and similar behaviour across different history windows, approximating observed experimental data in (Li & Dudman, 2013), that shows correlation below 0.2.

Figure 6 shows how the waiting time, relative to the reference blocks, changes as the subjects transition from high variance blocks to low variance blocks. Both algorithms show a gradual transition, although the SSP implementation has a sharper transition than the algebraic implementation. Likely this is due to the sharp discount of memories created by the low pass filter with $\gamma = 0.5$ (see Fig. 2).

Discussion

In this work we reproduce results from Li and Dudman (2013), Figure 3: waiting time has a linear dependence on the variance of the reward arrival distribution (see Figure 4). Importantly, our proposed method is based on a decision function that is agnostic to the type of distribution. Admittedly, the algebraic implementation of the algorithm uses a Gaussian distribution to compute it, but without losing generality in the decision making process. In the SSP implementation,
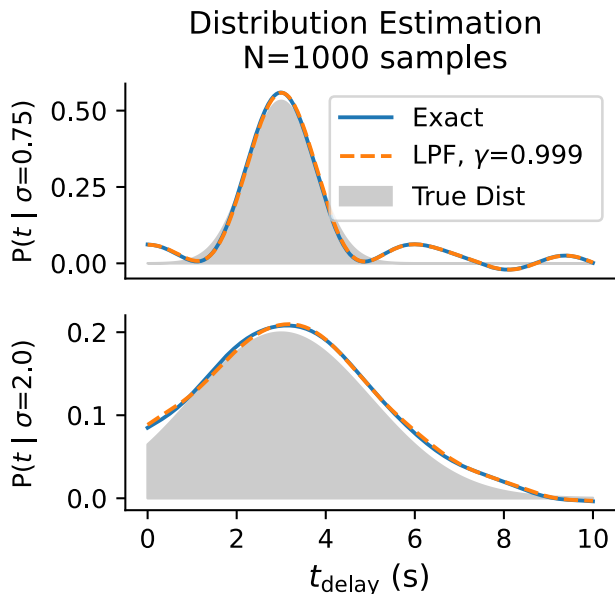
Given a set of observations, $\left\{t_{\text{delay},1}, \ldots, t_{\text{delay},N}\right\}$, the distribution over reward delay times can be embedded in a vector, μ_{delay}:

$$\mu_{\text{delay}} = \frac{1}{N}\sum_{i=1}^{N}\phi(t_{\text{delay},i}), \qquad (8)$$

but this representation requires keeping track of the number of observations, N. In order to make the estimation of the reward delay more realistic, compute μ_{delay} using a low-pass filter (LPF). LPFs can be implemented recurrently: $x' = \gamma x + (1-\gamma)z$, for state variable x, observation, z, and $\gamma \in (0,1)$. Our LPF on the vector μ_{delay} is defined on line 9 of Algorithm 2.

In Fig. 1 we show the effect of estimating a distribution using the exact definition from eq. 8 and by using a low-pass filter, with $\gamma = 0.999$, on 1000 observed delay times. For values of γ close to 1 and large sample sizes, the LPF approximate the empirical distribution. Estimating μ_{delay} with a LPF implicitly imposes an ordering effect on observed reward delays. Fig. 2 shows how observations are weighted as a function of the age of the observation in time steps.

Distribution Estimation
N=1000 samples

Figure 1: Estimated PDFs for t_{delay} using the SSP representation, for a window of 1000 samples for variance values of 750 and 2000 msec. We compute distributions using Eq. (8) (Exact), and using a low-pass filter with $\gamma = 0.999$ (LPF).

in contrast, we are using a non-parametric estimation of the distribution. Another important characteristic of the SSP algorithm is the effect of the discount factor, γ, on the memory for reward delays. In the work of Li and Dudman (2013), reward delay memory was modelled by a uniform sliding window of 10 observations, with perfect recall and perfect forgetting of anything outside of that window. We found that a discount factor of $\gamma = 0.5$ approximated mouse behaviour in the SSP implementation. Both models perform similarly relative to the data (as seen in Figure 4 A).

The main two differences between the algebraic and SSP implemented models are the method of estimation for elapsed time and how the reward time delay distribution is represented. Specifically, in Fig. 5, the SSP model's correlation does not grow monotonically with σ_{delay}, which is consistent with the experimental data (Li & Dudman, 2013, Figure S1A), with the reward distribution being learned from the elapsed time, as compared to the algebraic model that explicitly depends on previous history.

One advantage of our SSP model is that representations constructed in the HRR algebra are dimensionality-preserving. This means that more cues, either delivered in conjunction or with some temporal structure, will not require any additional memory – with the caveat that the dimensionality of the vector representation limits the representational capacity. Indeed, these representations have already been successfully employed in reinforcement learning tasks for biological and artificial agents (Bartlett et al., 2022a, 2022b, 2023). In contrast, the algebraic model, as well as the SET model, need to recruit new explicit memories or counters for
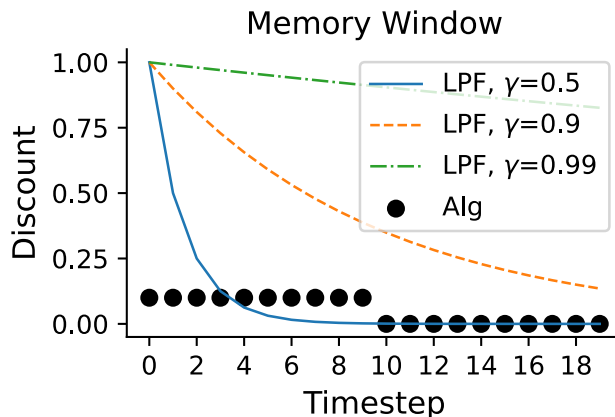
Memory Window

Figure 2: The discount is how much each sample contributes the estimated distribution parameters, as a function of sample age in timesteps. The algebraic implementation (Alg) has a perfect memory inside the history window. For smaller values of γ, the low pass filter (LPF) forgets samples faster.
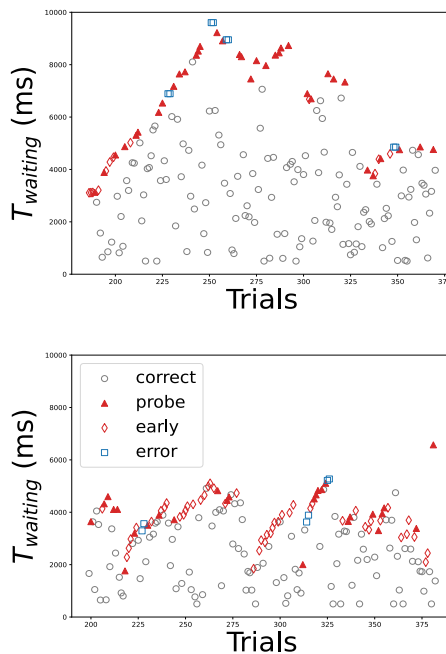

Figure 3: Waiting times for trials in an example $\sigma_{delay} = 2000ms$ block for the algebraic model (top) and SSP model (bottom). Legend shows types of trials consistent with those used by Li *et al.* (see Section *methods*).
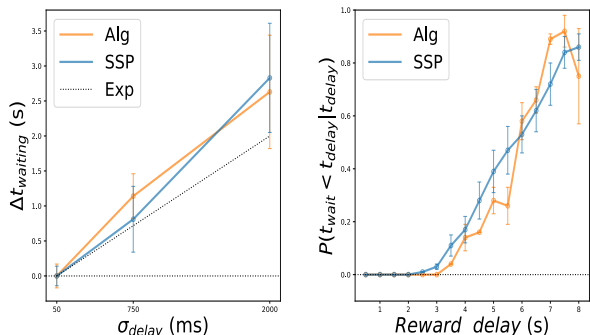
Figure 4: (Left panel) Average waiting time as a function of sigma (relative to $\sigma = 50$ msec) for probe trials in algebraic model (Alg) and SSP model. Dotted line represents observed trend in experimental data.(Right panel) Probability of leaving early in $\sigma = 2000$ msec blocks as a function of delay for rewarded trials.
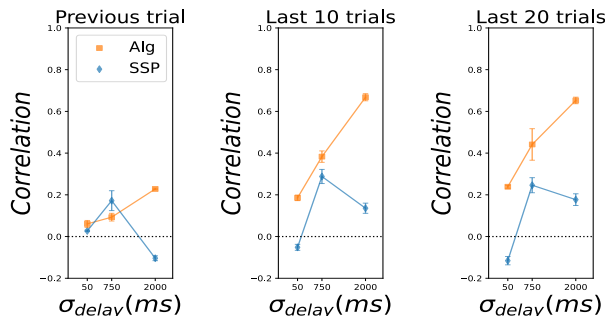

Figure 5: Correlation between waiting time in probe trials and delay times in previous rewarded trials with a history size of 1, 10 or 20 previous trials, for different σ_{delay}.
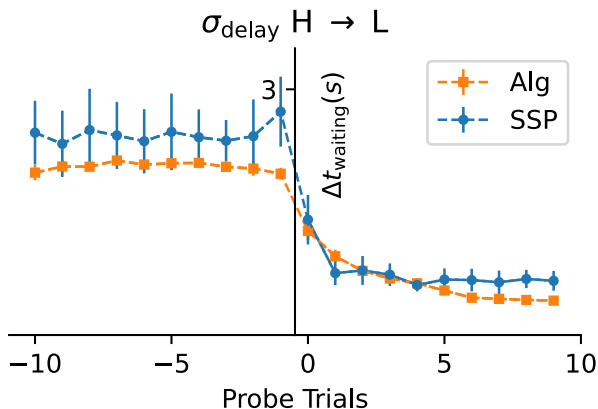

Figure 6: Here we show how $\Delta t_{\text{waiting}}$ changes when moving from high variance blocks to lower variance blocks. The results show average behaviour across subjects, error bars indicate SEM. Both algorithms display a smooth transition from the high to low variance blocks, but the algebraic implementation qualitatively fits better the smooth transitions observed in the experimental results of Li *et al.*.

each added relevant cue or factor determining outcomes.

Our proposed models resemble the SET model, in as much as we integrate temporal information and compare against a memory. In our SSP model, the recurrent connection (line 5 of Algorithm 2) is analogous to the pacemaker-accumulator of SET. However, the SET method samples previous reward timings from the memory, whereas we compare against the entire distribution. On the other hand, our algebraic model explicitly defines time step accumulation. An alternative approach is the LeT model, which uses a localist representation of successively activated states (neurons) to model time.

Like SET, our models do not take into account the reinforcement rate, although we could predict the reinforcement rate from the current memory state, $\phi(t_{\text{wait}})$, bridging the gap between the SET and LeT models in the SSP algorithm, through the use of a distributed representation of time. We suggest that our SSP model may encode something approximating the absolute reinforcement rates, like LeT. Since our low-pass filter allows for the extinction of memories of reward delivery, less frequently observed reward times will drop out of the memory. However, because our model is continually forgetting, it will not necessarily achieve a steady-state model of reinforcement rate.

Further, we conjecture that our model could be expanded to contextual-aware settings through the use of the VSA binding operation. The representation of time could be replaced with with a memory that integrates cues and selected actions, registered at the time the event took place, resulting in a μ_{delay} that represents distributions over the cue(s), selected action, and time of reward, *i.e.*, $\phi(t_{\text{wait}}) \circledast \phi(\text{cue}) \circledast \phi(a)$.

Our model assumes proficient mice that already know the task structure. In the future, we want to relax that assumption, and explore learning across blocks. This means learning also the lever-press decision.

Simen et al. (2011) propose the stochastic ramp and trigger (SRT) model of learning a timing task, resembling drift diffusion models. SRT integrates elapsed time using Poisson neurons with a linearly ramping firing rate. Our SSP algorithm integrates elapsed time with a recurrent network representing an SSP representation of time, $\phi(t_{\text{wait}})$. Our method resembles the sequential probability ratio test model of drift diffusion, implemented using SSPs (Furlong et al., 2023). However, where the previous SSP model integrates log evidence for a decision, here the log evidence of the integrated probability of the elapse time, given the reward structure. Regardless, our method shares the mechanism of thresholding log evidence to reach a decision with the SRT. By modelling this problem in the space of SSPs, we are hypothesizing about operations on the latent spaces represented by populations of neurons. Previous work has established that the interspike intervals of neural networks computing semantic pointers operations match the coefficient of variation found in biological systems (Komer & Eliasmith, 2016), so investigating if spiking neural implementations of our proposed model replicate firing activity observed in the SRT is warranted.

References

Bartlett, M., Simone, K., Dumont, N., Furlong, P. M., Eliasmith, C., Orchard, J., & Stewart, T. (2023). Improving reinforcement learning with biologically motivated continuous state representations. In *Proceedings of the 21st international conference on cognitive modeling.*

Bartlett, M., Stewart, T. C., & Orchard, J. (2022a). Biologically-based neural representations enable fast online shallow reinforcement learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Bartlett, M., Stewart, T. C., & Orchard, J. (2022b). Fast online reinforcement learning with biologically-based state representations. In *Proceedings of the 20th international conference on cognitive modeling.*

Dumont, N. S.-Y., & Eliasmith, C. (2020). Accurate representation for spatial cognition using grid cells. In *42nd annual meeting of the cognitive science society. toronto, on: Cognitive science society* (pp. 2367–2373).

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press.

Furlong, P. M., Bartlett, M., Stewart, T. C., & Eliasmith, C. (2023). Single neuron distribution modelling for anomaly detection and evidence integration. In *21st international conference on cognitive modelling (in press).*

Furlong, P. M., & Eliasmith, C. (2022). Fractional binding in vector symbolic architectures as quasi-probability statements. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Furlong, P. M., & Eliasmith, C. (2023). Modelling neural probabilistic computation using vector symbolic architectures. *Cognitive Neurodynamics*, 1–24.

Gibbon, J. (1977). Scalar expectancy theory and weber's law in animal timing. *Psychological review*, *84*(3), 279.

Glad, I. K., Hjort, N. L., & Ushakov, N. G. (2003). Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, *30*(2), 415–427.

Komer, B., & Eliasmith, C. (2016). A unified theoretical approach for biological cognition and learning. *Current Opinion in Behavioral Sciences*, *11*, 14–20.

Komer, B., Stewart, T. C., Voelker, A., & Eliasmith, C. (2019). A neural representation of continuous space using fractional binding. In *Cogsci* (pp. 2038–2043).

Lejeune, H., & Wearden, J. (2006). Scalar properties in animal timing: Conformity and violations. *Quarterly Journal of Experimental Psychology*, *59*(11), 1875–1908.

Li, Y., & Dudman, J. T. (2013). Mice infer probabilistic models for timing. *Proceedings of the National Academy of Sciences*, *110*(42), 17154–17159.

Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological review*, *104*(2), 241.

Machado, A., Malheiro, M. T., & Erlhagen, W. (2009). Learning to time: A perspective. *Journal of the experimental analysis of behavior*, *92*(3), 423–458.

MacLean, E. L., Mandalaywala, T. M., & Brannon, E. M. (2012). Variance-sensitive choice in lemurs: constancy trumps quantity. *Animal cognition*, *15*, 15–25.

Plate, T. A. (1992). Holographic recurrent networks. *Advances in neural information processing systems*, *5*.

Plate, T. A. (2003). *Holographic reduced representation: Distributed representation for cognitive structures*. Stanford, CA, USA: CSLI Publications.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*(11), 2745–2752.

Simen, P., Balci, F., deSouza, L., Cohen, J. D., & Holmes, P. (2011). A model of interval timing by neural integration. *Journal of Neuroscience*, *31*(25), 9238–9253.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Yi, L. (2007). Applications of timing theories to a peak procedure. *Behavioural Processes*, *75*(2), 188–198.

How to Provide a Dynamic Cognitive Person Model of a Human Collaboration Partner to a Pepper Robot

Alexander Werk (alexander.werk@campus.tu-berlin.de)
Institute of Psychology and Ergonomics, Berlin Institute of Technology, Marchstraße 23 10587 Berlin, Germany

Sina Scholz (s.scholz@uni-luebeck.de), Thomas Sievers (t.sievers@uni-luebeck.de),
Nele Russwinkel (nele.russwinkel@uni-luebeck.de)
Institute of Information Systems, University of Lübeck, Ratzeburger Allee 160 23562 Lübeck, Germany

Abstract

For successful and trustful human-robot interaction, the challenge is to provide the robot with information so it can adapt dynamically to humans and changing situations. Cognitive architectures such as ACT-R provide valuable capabilities to address this challenge. This paper demonstrates how cognitive architectures can be used to improve human-robot interaction. First, this paper illustrates how mental representations can be built up in order to anticipate the partner and the situation and collaborate adaptively. Second, it is shown how a model can be easily integrated into a robot. Finally, this paper provides an example of how emotion recognition can be used to adapt the interaction accordingly by utilizing perceived changes in the real world. As a result, the paper presents instructions, concepts and use cases for implementing the various aspects. The paper encourages further research on how cognitive architectures can address challenges in human-aware AI.

Keywords: Anticipation, Person Models, Human-Robot Interaction, Emotion Recognition, Social Robots, Adaptive Behavior, Cognitive Models, ACT-R, Human-Aware AI

Introduction

Robots interacting with humans and solving tasks together with a human partner need to have some kind of model of the world, the situation, the task to be solved and the person it is interacting with. So far, there are robots that have impressive capabilities such as navigating, interacting with objects and showing social interaction (e.g. Rossi et al. (2019)). However, there is still a lack of robots' capabilities needed to adapt to changes in the state of the human partner or of circumstances due the developing situation. According to Kambhampati (2019), research in Artificial Intelligence must pay more attention to aspects of intelligence that help humans work with each other – including social intelligence. Kambhampati (2019) introduced the term human-aware AI systems as "goal directed autonomous systems that are capable of effectively interacting, collaborating, and teaming with humans". The challenges in designing such human-aware AI systems, include "modeling the mental states of humans in the loop, recognizing their desires and intentions, providing proactive support, exhibiting explicable behavior, giving cogent explanations on demand, and engendering trust" (Kambhampati, 2019). The approach to provide a robot with cognitive models, or to use a cognitive architecture for this is close at hand. Kurup and Lebiere (2012) made several suggestions how cognitive architectures can offer support e.g. by offering strong models for interacting with dynamic environ-

ments. The authors mention several points that require high-level cognition in human-robotic interaction (HRI). Two of these will be addressed here: (1) flexible, adaptive, dynamic and real-time behavior and (2) interacting with humans in a natural way. Regarding (1) robust real-world behavior cannot be pre-programmed. It requires flexibility and building up of representations that can be updated and implies the ability to understand the current situation and how it developed. With regard to (2), the focus is on the ability to understand the human partners actions and intentions and react appropriately. These two points still haven't been addressed sufficiently for HRI. In this paper, we want to introduce a concept of person models in combination with situation representation that can be used for HRI and to argue how and why cognitive architectures offer valuable approaches to solve existing problems in dealing with real world and collaboration challenges for robots. The paper will be structured in three main parts:

First, the core concept of person and situation models for a collaborative task in human-human interaction is introduced and then transferred to HRI and towards an cognitive model approach. **Second**, we show how cognitive models can be integrated with a robot. For this, cognitive models must be structured differently, i.e. as model tracing approaches, which we call anticipatory models (e.g. (Klaproth et al., 2020) or (Hao, Russwinkel, Haeufle, & Beckerle, 2023)). This kind of models continuously takes input information from the real world and builds representations by observing and interpreting the information in relation to the stored general knowledge. This can be task knowledge, general information about human partners, or data about environmental changes. There have been several approaches to enable a robot with cognitive functions by connecting it with ACT-R. One approach ACT-R/E (Trafton et al., 2013), was placing an additional constraint on cognition namely that cognition occurs within a physical body that must navigate in real surroundings, as well as perceive the world and manipulate objects with a complex extension of the cognitive architecture ACT-R. In this paper we show how to connect ACT-R in a simple way with robots such as Pepper via the goal buffer. **Third**, as an example for information perceived from the environment, recognition of emotional states is taken. In a use case, the importance of emotion recognition and interpretation for adaptive HRI is demonstrated (e.g. (Weidemann & Rußwinkel, 2021). Based on the work of psychologist James Russell (J. A. Russell,

2009), a simplified concept of basic emotions like neutral, content, joyful, sad and angry, which is already implemented in the Pepper robot, was used to enable adapted behaviour of the robot towards humans.

The Person Model Theory

The Person Model Theory (PMT; Newen, 2015) assumes that there are three types of models for analyzing other people or situations: (i) **person models**, (ii) **self models** and (iii) **situation models**. The **person models** are representations that summarize certain characteristics of another person or a group of people. Equivalently, the **self model** contains the same information as the **person models**, only in relation to the person him/herself. In contrast to these two types of models, **situation models** contain external information about the environment and further contextual information that goes beyond the knowledge from the **person models** (e.g. A person comes to the town hall. People coming to the town hall usually have a concern. Therefore, the person has a concern that he/she wants to address today.). Furthermore, Newen (2015) postulates different types of understanding, on the one hand implicit, intuitive understanding and on the other hand inference-based understanding. This means that the **person models** or **self models** can also be further subdivided into implicit person schemas/self schemas and explicit person images/self images. The former include, for example, information about facial expressions, emotions, gestures, moving patterns and body posture of another person or the person him/herself (e.g. The other person is joyful.). Person images/self-images, in contrast, contain information such as names, descriptions, biographies and visual images (e.g. The other person's name is John.). Furthermore, one claim of PMT is that children are already able to create models about other people, themselves and the situation in this way, and that these models are available early on in human development (Newen, 2015).

Due to this extensive possibility of collecting and storing knowledge in a targeted manner and its early availability in human development, this approach seems to hold not too much complexity and could be a good approach to be integrated into social robots in order to adapt dynamically to a partner.

Cognitive Modelling with ACT-R

To integrate the PMT into a robot, the ACT-R cognitive architecture was chosen. ACT-R is based on cognitive psychology research, which means that ACT-R models follow certain rules and are subject to certain constraints. These include the different modules that are associated with neural correlates through imaging studies (Anderson et al., 2004). The exchange of information between the individual modules takes place via so-called buffers, through which the modules are connected to each other. The knowledge units containing the information can be requested from designated buffers. These so-called chunks can also be changed in the buffers. They

comprise a chunk name, chunk type and slots that can be customized to define the number and content. These mechanics are assigned to the symbolic processing of ACT-R. However, ACT-R not only has symbolic processing, but also subsymbolic processing, which is why ACT-R is also considered a hybrid architecture (Kotseruba & Tsotsos, 2018). Subsymbolic processing offers promising opportunities, especially for further development of the theoretical approach and the integration of this approach into complex ACT-R models, e.g. in terms of utility learning or influencing production rule selection by adjusting the utility. In the context of this paper and the presentation of the theoretical approach, the focus in this paper will be on symbolic processing for the sake of simplicity.

The nature of knowledge representation in the form of chunks and information exchange in the form of productions on the one hand, and the potential for refinement of the theoretical approach through subsymbolic processing on the other, make the ACT-R architecture a suitable choice for modeling the components from PMT.

Combining Person Model Theory and ACT-R

To implement the PMT in an ACT-R model, three requirements must be met: (1) The model must be able to incorporate information from the environment provided by the robot. (2) It must use the environmental information to keep both the **person model** and the **situation model** up to date. (3) Finally, the model must send an appropriate output back to the robot so that it can adapt its actions to dynamically changing environmental conditions.

The primary task of the model is to build and update the **person models** and **situation models**. Both contain dynamically changing information units on the one hand and information units that do not change during the interaction on the other. In the **person models**, the dynamically changing information units include facial expressions, movement patterns or gestures, which can constantly change depending on the course of the interaction and the current mood. In the **situation models**, these information units include details on the current status of the interaction, i.e. the current action, the current subgoal and the current phase. On the side of the more static information units, information such as age, gender, name, and biographical facts must be included in the **person models**. Whereas information such as the type of interaction (e.g. conversation or collaboration in complex tasks) must be included in the **situation models**. For a simple use case in which only one person interacts with the robot at a time (e.g. A person wants to apply for a new passport.), the **person model** and the **situation model** can be combined in one chunk, which is referred to below as the **model chunk**. Since the information in the **model chunk** is constantly updated during the interaction, the imaginal buffer was selected to hold this chunk so that the **model chunk** can be permanently maintained and adapted at any time. At the beginning of each interaction, the imaginal buffer was free. When the robot transmitted the information to the ACT-R model that

the interaction had started, a request was sent to the imaginal buffer to create a new **model chunk**. The **model chunk** contained several slots for the personal information - name of the person, age, gender, emotion, language, simple language - and several slots for the situation-related information - the customer's intention, the current sub-goal, the current action, the next action, the current phase, and the next phase. Whenever an input chunk with personal or situation-related information was transmitted to the ACT-R model, the corresponding slot in the **model chunk** was filled or updated. When the **model chunk** was created at the start of the interaction, only the slots for the current action, current phase and the current subgoal were filled with values. Figure 1 shows the schematic structure of a **model chunk** for a use case with one interaction partner.

It should be noted that in use cases where several people interact with the robot or interaction partners change frequently, the **person model** and the **situation model** should be stored in separate chunks and edited in the imaginal buffer as required.

```
model chunk
        isa             model
        person_name     (e.g. John Doe)
        gender          (m/f/d)
        age             (e.g. 24)
        language        (e.g. English)
        simplelang      (Yes/No)
        emotion         (e.g. Joyful)
        intention       (e.g. New Passport)
        current_action  (e.g. Transfering)
        next_action     (e.g. Farewell)
        current_phase   (e.g. Transfer)
```

Figure 1: Schematic structure of a model chunk

This multitude of information in the slots of the **model chunk** offers a variety of possible combinations and enables the model to send a dynamic output back to the robot that is individually adapted to the person and situation. In addition, the complexity of the model can be flexibly adjusted by selecting certain characteristics to be taken into account from the **model chunk**. As described above, the output is transmitted to the robot via the *pepper_out* slot of the goal chunk, in which a custom value can be entered.

Connect ACT-R to a Pepper Robot

The use of cognitive architectures is promising for achieving more human-like reactions and behavior in robots. Their formalized models can be used to further refine a comprehensive theory of cognition in order to provide a common ground for working towards a specific goal, responding flexibly to actions of the partner and developing a situational understanding for adequate reactions.

Humanoid Robot Pepper

The social humanoid robot Pepper developed by Aldebaran (Aldebaran, United Robotics Group and Softbank Robotics, n.d.), as shown in Figure 5, is 120 centimeters tall and optimized for human interaction. It is able to engage with people through conversation, gestures and its touch screen. Pepper can focus on, identify, and recognize people. Speech recognition and dialog is available in 15 languages. Beyond, Pepper manages to perceive basic human emotions. The robot features an open and fully programmable platform so that developers can program their own applications to run on Pepper.

Since research has generally shown that trust is the basis for successful communication tasks and trust in robots is increased by anthropomorphism, a humanoid social robot like Pepper is a good choice for social interaction and the provision of services when dealing with customers. A human face, the possibility of human-like expressions and body language and the use of voice are seen as beneficial for the trust of customers in the robot (Fink, 2012). The Pepper robot is already being used in many HRI projects and has also been tested in real production use.

System Setup for ACT-R and the Robot

We used the *standalone* version of ACT-R, i.e. the application provided at `https://act-r.psy.cmu.edu/`. To establish a remote connection from the robot to ACT-R, the remote interface – the **dispatcher** – was used, which is implemented by a central command server. The ACT-R core software connected to this dispatcher to provide access to its commands, and the dispatcher accepted TCP/IP socket connections that allowed clients to access these commands and provide their own commands for use. The commands available via the dispatcher can be used wherever a Lisp function was formerly required. The use of the dispatcher is described in detail in the ACT-R RPC Interface Documentation (remote.pdf in the "docs" directory of the ACT-R installation).

By default, the standalone version forces the dispatcher to use the localhost IP address of the computer on which it is running for connections instead of an external IP address. This means that only programs on the same computer can establish a connection, and once ACT-R has been started, this can no longer be changed. To disable this function, the file force-local.lisp must be removed from the ACT-R/patches directory before the application is executed. Then it will use the machine's real IP address for the dispatcher's connections, and the setting *allow-external-connections* in the model file will let other machines connect. Another option is to place the model file in the ACT-R/user-loads directory. External connections are then always permitted. The address and port used by the dispatcher is displayed at the top of the ACT-R terminal window. This information must be used on the remote computer for the connection.

The client application we developed for Pepper contained a program section for the remote connection to the dispatcher. A very basic example of a general setup of such a connec-

tion can be found in the "examples/connections" directory of the ACT-R installation for various programming languages. This client connection was used to start and control an ACT-R model that represented the cognitive processes for controlling human-robot interaction. The client was able to interact directly with the model by calling commands. The *run-full-time* command, together with a number of seconds, started and ran the model for the specified time. The *evaluate* method was used to evaluate commands from the dispatcher. It required the name of the command to evaluate. Figure 2 depicts the complete system setup including the interaction of its subcomponents.
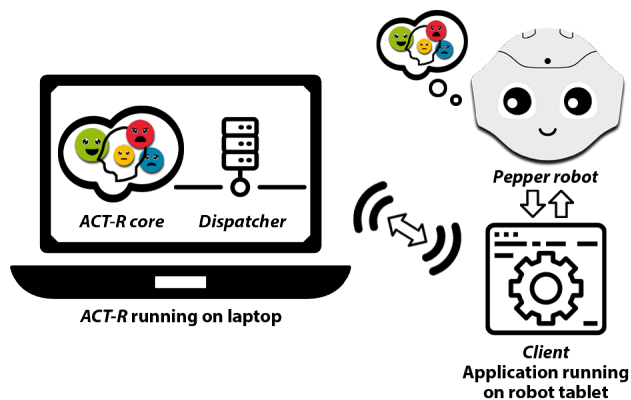
Figure 2: System Setup including the interaction of its subcomponents.

Our ACT-R model created in Lisp used a goal slot *pepper_out* for sending commands to the client application using ACT-R productions. This goal slot was evaluated via a permanently running while loop using the *buffer-slot-value* command that got the value of a slot from the chunk in a buffer of the current model. The *buffer-slot-value* was sent as a string in JSON format via the TCP/IP socket stream. A unique ID was assigned to each evaluation command to identify the correct part of the data in the stream received by the socket. The permanent evaluation of the content of the goal slot *pepper_out* in the client application was used to create special commands for the robot depending on this slot content, e.g. to execute a certain animation or to make a corresponding utterance.

To illustrate the syntax, the following lines show an example of using the *evaluate* method to retrieve a goal slot as a control signal from the model. This was done using the *buffer-slot-value* command in a while loop and a production in the Lisp code of the ACT-R model using a goal slot *pepper_out* for sending such a signal to the client application:

```
while (true) {
 ...
 print("{method:evaluate, params:
 [buffer-slot-value, nil, goal, pepper_out], id:10}")
 ...
}
```

```
(p checking-intention
   =goal>
      isa goal
      next_a checking-intention
==>
   =goal>
      next_a clearing-pepper-out
      pepper_out pepper-checks-intention
)
```

To transmit information from the client application on the robot to the ACT-R model, the client used the *overwrite-buffer-chunk* command to copy a chunk into the goal buffer. Therefore, the model had predefined goal chunks in its declarative memory. If a predefined chunk matched the chunk from the client, all information from this predefined chunk were placed in the buffer and could be used to trigger a production in the model.

Combining Real World Information with ACT-R

The aim was to transfer the basic emotions identified by Pepper into an ACT-R model. The input from the robot was received by the model by calling predefined chunks in the goal buffer. The goal buffer was chosen to be a good candidate as transfer buffer for a first simple approach. On the long run a specific buffer holding information that is transferred from Pepper would be necessary. Figure 3 shows the schematic integration of the ACT-R architecture into the robot.
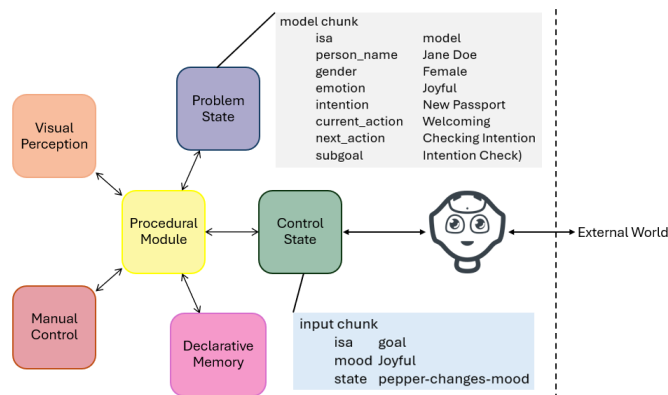

Figure 3: The ACT-R cognitive architecture according to Borst and Anderson (2017) - edited by the authors. The Pepper robot transmits information from the external world as an input chunk to the goal buffer of the ACT-R model. From there, the information can be transferred to other chunks, such as the model chunk in the imaginal buffer.

For transmitting a recognized emotion the *overwrite-buffer-chunk* command was used to trigger the appropriate productions of the ACT-R model. Predefined goal chunks in the declarative memory of the model made it possible to control the productions depending on the transmitted emotion values. Examples of such goal chunks, which were prepared in the Lisp code of the ACT-R model, and an example production that filled a *pepper_out* goal slot with a value that

was evaluated in the client application of the robot, can be found in the following lines:

```
(add-dm
 (mood-content-chunk isa goal mood content state pepper-changes-mood)
 (mood-joyful-chunk isa goal mood joyful state pepper-changes-mood)
 (mood-sad-chunk isa goal mood sad state pepper-changes-mood)
 (mood-angry-chunk isa goal mood angry state pepper-changes-mood)
)

(p pepper-content
   =goal>
      isa goal
      mood content
      state pepper-changes-mood
==>
   =goal>
      next_a clearing-pepper-out
      pepper_out pepper-content
      state pepper-changed-mood
)
```

The robot's responses controlled by the client application were influenced in this way. Depending on the goal slot value, different dialogues or responses were triggered on the robot side. For anticipatory models that anticipate events in the real world and feed this into a model (e.g. (Klaproth et al., 2020) or (Hao et al., 2023)), the main idea is to not initiate all perceptual processes as usually done in modelling. Instead, the main processes work on mental representations that are built up and hold the momentary understanding of a situation. Productions and retrieved chunks depend on the situation understanding. Therefore, anticipatory models continuously take input information from the real world and build up representations such as a person model in this example.

Emotion recognition in Human-Robot Interaction

By integrating emotion recognition in HRI, robots can respond more effectively to the needs of people, leading to an improved collaboration and a more pleasant experience for humans. This ability of the robot to recognize emotional states and behave appropriately is particularly significant and an important basis for building trust in HRI (Jessup & Schneider, 2021). To recognize and categorize emotions, there are various technical methods that can be used by robots: *Facial Recognition*, *Voice Recognition*, *Body Movement and Gesture*, *Physiological Signal Processing* and more established approaches like *Multi-modal Approaches*.

Pepper's perceptive abilities

The Pepper robot was developed with a focus on social interactions with humans. It is equipped with the ability to perceive its environment, including the emotional states of human interaction partners. Pepper can perceive numerous personal features, such as *age, gender, smile state, mood (pleasure state), excitement state, attention state, engagement intention state* (Softbank Robotics, n.d.). Particularly important for emotion recognition are the *mood/pleasure state* and the *excitement state*. The *pleasure state* is based on facial features, touch and speech semantics and can have the values *positive*, *neutral* or *negative*. The *excitement state* is based on

the interaction partner's voice and can have the values *calm* or *exited*. Based on the work of James Russell (J. A. Russell, 2009), a transformation matrix shown in Table 1 was used for the conversion of these states into the emotions *neutral*, *content*, *joyful*, *sad* and *angry*.

Table 1: Transformation matrix to get the basic emotions

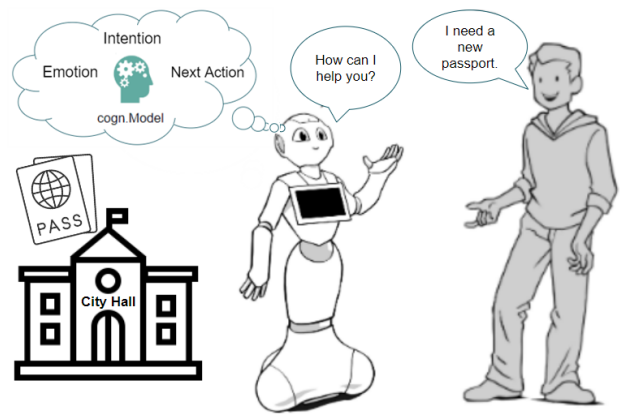| ExitementState	PleasureState		
	Positive	Neutral	Negative
Calm	Content	Neutral	Sad
Exited	Joyful	Neutral	Angry

Figure 4: Pepper using person model (Softbank Robotics / edited by the authors, 2024)

Adapted behaviour for successful task completion

To test our approach, a scenario-based study was conducted with the Pepper robot. As part of the scenario, the robot worked in a town hall and was tasked with helping customers apply for a new passport. The Pepper robot helped the human to complete this task as shown in Figure 5. Pepper took the perceived emotions of the person into account and thus showed situationally appropriate behavior. The approach was tested using three different scenarios:

- problem-free (task can be solved without any problems)

- complication (task with complication, but can be solved)

- problem (task can not be solved)

As in real life, emotional states can change during the scenarios. For example, the **complication scenario** can evoke different emotions in people and lead to disturbances in coping with the task. Depending on individual dispositions, humans deal with such situations differently. In this scenario, the person did not have the bio-metric images with them, even though they were needed to apply for a passport. When the person realizes that she does not have the passport photos
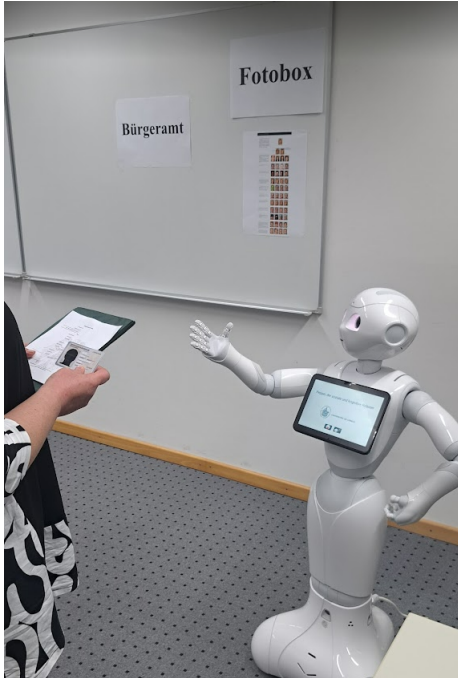
Figure 5: Pepper in a collaborative task with a person that wants to check if necessary documents are complete

with her, she can conclude that the task can no longer be completed successfully. This could possibly lead to **anger** (e.g. Being frustrated about forgetting passport photos) or **sadness** (e.g. Worrying that one won't be able to go on holiday).

An adequate communication style was developed for Pepper that corresponded to the new emotional state of the person and the situation. The robot's adjusted behavior was designed to support the achievement of the goal and the completion of the task. As an example, we adjusted the response regarding the emotional state of a person who was perceived as being *sad* with the following phrases:

sad: *I _understand_. So, you don't have any passport photos with you. _It can happen to anyone_. However, it's _not a big problem_. You can go next door to the photo booth to take bio-metric passport photos and then _come back to me_.*

The underlined text was meant to support the person in this situation by showing that Pepper understood the situation, reassuring and encouraging the person. In addition to the customized communication, the gestures of Pepper were also adapted depending on the emotional states of the person. For the sake of simplicity, this article is limited to the presentation of verbal response behaviour.

Discussion

This paper shows how to provide a robot with a dynamic mental representation to enable an adaptive, human-like social interaction. Cognitive architectures such as ACT-R are relevant for designing agents that can act flexibly in the real

world, e.g. reacting to new situations and adapting to the task, environment and interaction partners. The aim of the paper was to highlight core concepts and provide guidelines on how this can be achieved and implemented. Overall three aspects were presented: **First**, the core concept of person and situation models necessary for acting adaptable to a partner in a collaborative task was introduced. In our example, the person and situation models were implemented as one chunk. For more complex tasks and interaction situations, it might be necessary to hold the information separately. **Second**, we showed how cognitive models can be integrated into a robot. One aspect of this is how anticipatory models can be designed in order to continuously incorporate information from the real world and build dynamic representations which form the basis for further action decisions. Furthermore, a simple approach is introduced on how to realize the information flow between the cognitive architecture and the robot. **Third**, as an example of external information perceived by the robot, changes in the emotional state of the human partner were used. Changes in the mental state itself, but also the change of an emotion in a specific situation, can provide relevant information for adapting appropriately to a partner. One of the biggest challenges of emotion recognition in HRI is the reliability of the AI systems that are used. Whether emotion recognition without context actually allows direct conclusions about a person's inner state has been criticized e.g. (Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019) (Weidemann & Rußwinkel, 2021). The actual emotional state of a person depends on the situation or context and the individual and therefore cannot be reliably determined using data-based approaches alone. For human compatible AI according to S. J. Russell (2019), it is necessary that the system is uncertain about the human state and should therefore collect further evidence from human behavior. One of the **first future steps** is to consider state and emotion changes more closely to investigate interactions between humans and machines that have human-like cognitive abilities with the help of appropriate models. A **second future step** should be to include more flexible task knowledge that is chosen according to the state of the representations. The evaluation of such a system should be a **third future step**. However, the adaptable nature of the interaction might need long and complex interactions to be perceived. Therefore, the example needs to be further expanded and then evaluated according to the perceived transparency and perceived naturalness of the interaction. Regarding the challenges with emotion recognition, in a **fourth future step**, the emotional state should also be derived from the evolving situation in order to reconcile multiple sources of information for a more reliable and human-like interaction.

References

Aldebaran, United Robotics Group and Softbank Robotics. (n.d.). *Pepper* (Tech. Rep.). Retrieved 2024-02-20, from `https://www.aldebaran.com/en/pepper`

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1-68. doi: 10.1177/1529100619832930

Borst, J. P., & Anderson, J. R. (2017). A step-by-step tutorial on using the cognitive architecture act-r in combination with fmri data. *Journal of Mathematical Psychology*, *76*, 94–103.

Fink, J. (2012). *Anthropomorphism and human likeness in the design of robots and human-robot interaction*. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-34103-8_20

Hao, C., Russwinkel, N., Haeufle, D. F., & Beckerle, P. (2023). A commentary on towards autonomous artificial agents with an active self: Modeling sense of control in situated action. *Cognitive Systems Research*, *79*, 1-3. Retrieved from https://www.sciencedirect.com/science/article/pii/S1389041722001085 doi: https://doi.org/10.1016/j.cogsys.2022.12.006

Jessup, S. A., & Schneider, T. R. (2021). Chapter 22 - the role of emotions in human-robot interactions. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (p. 515-530). Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780128194720000228 doi: https://doi.org/10.1016/B978-0-12-819472-0.00022-8

Kambhampati, S. (2019). Challenges of human-aware AI systems. *CoRR*, *abs/1910.07089*. Retrieved from http://arxiv.org/abs/1910.07089

Klaproth, O. W., Halbrügge, M., Krol, L. R., Vernaleken, C., Zander, T. O., & Russwinkel, N. (2020). A neuroadaptive cognitive model for dealing with uncertainty in tracing pilots' cognitive state. *Topics in cognitive science*, *12*(3), 1012–1029.

Kotseruba, I., & Tsotsos, J. K. (2018). A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications. *arXiv preprint arXiv:1610.08602v3*.

Kurup, U., & Lebiere, C. (2012). What can cognitive architectures do for robotics. In *Biologically inspired cognitive architectures*. Retrieved from https://api.semanticscholar.org/CorpusID:62720059

Newen, A. (2015). Understanding others - the person model theory. *Open MIND*. doi: 10.15502/9783958570320

Rossi, A., Cruz Maya, A., Dautenhahn, K., Koay, K., Walters, M., & Pandey, A. K. (2019, 06). Investigating the effects of social interactive behaviours of a robot on people's trust during a navigation task. In (p. 349-361). doi: 10.1007/978-3-030-23807-0_29

Russell, J. A. (2009). Emotion, core affect, and psychological construction. *Cognition and Emotion*, *23*(7), 1259-1283. doi: 10.1080/02699930902809375

Russell, S. J. (2019). *Human compatible : artificial intelligence and the problem of control*. London: Allen Lane/Penguin Books.

Softbank Robotics. (n.d.). *Perceptions human characteristics* (Tech. Rep.). United Robotics Group (URG). Retrieved 2024-02-22, from https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/ch4_api/perception/reference/human.html

Softbank Robotics / edited by the authors. (2024). *Figure: Pepper and human in comic style (softbank) and pass and citi-hall pictograms* (Tech. Rep.). Softbank Robotics / The Noun Project. Retrieved 2023-06-28, from https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/ch6_ux/chap2.htmlandhttps://static.thenounproject.com/png/1969047-200.png

Trafton, G., Hiatt, L., Harrison, A., Tanborello, F., Khemlani, S., & Schultz, A. (2013, 03). Act-r/e: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, *2*, 30-55. doi: 10.5898/JHRI.2.1.Trafton

Weidemann, A., & Rußwinkel, N. (2021). The role of frustration in human–robot interaction–what is needed for a successful collaboration? *Frontiers in psychology*, *12*, 707.

"I *Knew* It!" Model-Based Dissociation of Prior Knowledge Confounds in Memory Assessments

Alyssa Williams (arw698@msstate.edu)
Department of Psychology, Mississippi State University, Starkville, MS 39759 USA

Holly Sue Hake (hakehs@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Andrea Stocco (stocco@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Abstract

Computational modeling is a powerful approach for discerning individual differences in memory function. The model-based assessments discussed in this paper rely on estimating an individual's rate of memory decay– a stable and idiographic parameter that can be captured by the model. However, this paper aims to demonstrate prior knowledge as a confounding factor in these model-based assessments and seeks to parse out the error using Maximum Likelihood Estimations. The metric of individualized memory performance, termed *Speed of Forgetting*, was significantly lower for facts known beforehand. Still, these facts were identified with 81% accuracy by recovered base-level activation estimations blind to the ground-truth data. A proposal for future model-based assessments to account for prior knowledge is discussed.

Keywords: ACT-R, Cognitive Neuroscience, Computational Modeling, Memory, Prior Knowledge

Introduction

Reliable assessment of memory function is essential to conducting research on memory processes, understanding memory-related disorders, and developing new therapeutic interventions. Memory function is typically assessed through performance in response to memory probes. However, these responses not only reflect the underlying accessibility of memory but also other confounding factors.

Among these confounds, prior knowledge–i.e., the possibility that the participant might already know the answer–is perhaps the most significant. Researchers have attempted to address the issue of prior knowledge by employing novel artificial or abstract stimuli. However, these stimuli are often challenging to encode initially, leading to an underestimation of memory function and rendering them unsuitable for clinical use (Brady et al., 2008). Alternatively, memory researchers have used paired-associates to examine how novel associations between familiar items (e.g. "fireman" and "slug") are learned (Anderson, 1974). Nonetheless, random associations are susceptible to semantic congruency effects if the stimuli are not meticulously chosen. For example, "fish" - "sea" would have higher congruence than "zebra" - "sea" (van Kesteren et al., 2012). Analyses from Sense and van Rijn (2022) confirm that prior knowledge should not be neglected and used subject-specific grades as a proxy to control for prior knowledge. Overall, it can be useful for researchers to identify and mitigate the effects of prior knowledge rather than continually designing new stimuli.

In this paper, we demonstrate the feasibility of such an approach in the absence of a current proxy. This method relies on model-based assessments of memory functions, in which a participant's long-term memory function is delineated as a parameter of a model fitted to their data. We illustrate that the impact of prior knowledge can be conceptualized as an additional item-level parameter in the model. Moreover, we establish that through maximum-likelihood parameter recovery procedures, it is possible to accurately discern the extent to which a specific memory item was previously known.

Model-Based Assessments

Central to this paper is the use of *model-based* assessments, which are memory function evaluations predicated on the value of a parameter within a computational memory model fitted to the data. Model-based assessments serve as a type of "cognitive twin", reflecting an individual's cognitive processes (Somers et al., 2020). Significantly, their applications have recently been integrated into clinical settings where appraisal of memory ability is critical, such as in evaluating mild cognitive impairment and dementia (Nasreddine et al., 2005). In this paper, we will use the memory model originally proposed by Anderson and Schooler (1991) and later incorporated as part of the ACT-R cognitive architecture. This model can be viewed as a computational implementation of the Multiple Trace Theory (Nadel et al., 2000). According to this theory, a "memory" is the accumulation of multiple episodic traces during which the same information has been presented.

In essence, the activation of a memory, or odds of its retrieval, increases with each trace but gradually decays over time. For example, when the Spanish word "manzana" for "apple" is initially learned at time 0, the activation of that fact experiences a spike. Subsequently, that fact would gradually be forgotten following a power curve until the word's meaning is encountered again, perhaps at time 20 (Figure 1). Thus, as the meaning of "manzana" is learned, the activation level increases, and the rate of decay slows.
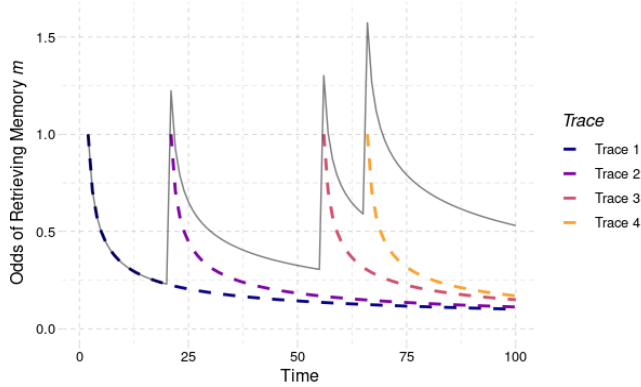
1

Figure 1: Hypothetical time course of the activation of a memory (gray solid line) made of four different traces (colored dashed lines) encoded at four different times.

More formally, a single memory m corresponds to a set of decaying traces encoded at different times, indicated as $t(1)$, $t(2)$, … $t(N)$. The odds of retrieving each trace decay are a power function of time, as dictated by the power law of forgetting (Newell & Rosenbloom, 1981). In other words, the odds of retrieving a memory m at a given time t are proportional to its activation $A(m, t)$, which is the logarithmic sum of all of its decaying traces i over time:

$$A(m, t) = \log \sum_i (t - t(i))^{-d(i)} \quad (1)$$

Where $(t - t(i))$ is the time since the creation of the i-th trace, and $d(i)$ is a trace-specific decay rate. The trace's decay rate, in turn, is a function of the memory's residual activation at the time of the i-th trace's creation, $t = t(i)$, plus a constant forgetting momentum φ:

$$d(i) = e^{A(m, t = t(i))} + \varphi \quad (2)$$

The parameter φ represents the initial decay rate of a newly formed memory and determines the decay rates of all subsequent traces. This parameter is also known as the *Speed of Forgetting* (SoF) and accounts for the spacing effect and recency effect in memory retention (Cepeda et al., 2008). Experimental research has shown that this parameter is a latent characteristic of an individual, is stable over time, and is reliable across experimental sessions and stimuli (Hake et al., 2023; Sense et al., 2016; Zhou et al., 2021).

Using this underlying model to estimate the *SoF* parameter, this model based assessment provides a fast, easy, and reliable way to assess an individual's memory ability. Due to the applicability of model-based assessments to clinical memory impairment, improving model accuracy is of utmost importance. Within this computational framework, accounting for prior knowledge within the model is a significant step toward this goal.

Representing Prior Knowledge

The model provides an intuitive way to represent prior knowledge computationally. Generally, the activation of a

previously known item results from the combined contribution of n experimentally observed traces and m unobserved previous traces, which are inherently inaccessible. To simplify computational representation, we make a key assumption.

In most cases, prior knowledge has been acquired well before the experiment begins. This means that the effect of temporal decay is negligible within the context of a single experimental session. This is illustrated over the course of 12 months in Figure 2, which plots the activations of three hypothetical memories that have accumulated 1, 10, or 100 traces over the first month. While the number of associated traces has a sizable effect on their residual activation, the effect of forgetting becomes negligible over time. Thus, we can assume that the effect of prior knowledge is an essential constant over time in the course of our experiment. For this reason, we will simply write that the "true" activation $A'(m)$ of a memory m is the sum of the contribution of the traces accumulated over an experimental session plus a memory-specific constant K_m:

$$A'(m, t) = \log \sum_i (t - t(i))^{-d(i)} + K_m$$
$$= A(m, t) + K_m \quad (3)$$

In ACT-R terms, the parameter K_m represents a memory-specific *base-level constant* (BLC) that summarizes the previous history of a memory before an experiment takes place. While typically ignored, the presence of such a constant dramatically affects the estimates of other model parameters from experimental data. However, as the remainder of this paper will show, because the distortions introduced by BLC can be modeled as well (as in Equation 3), their contributions can be automatically estimated and corrected.

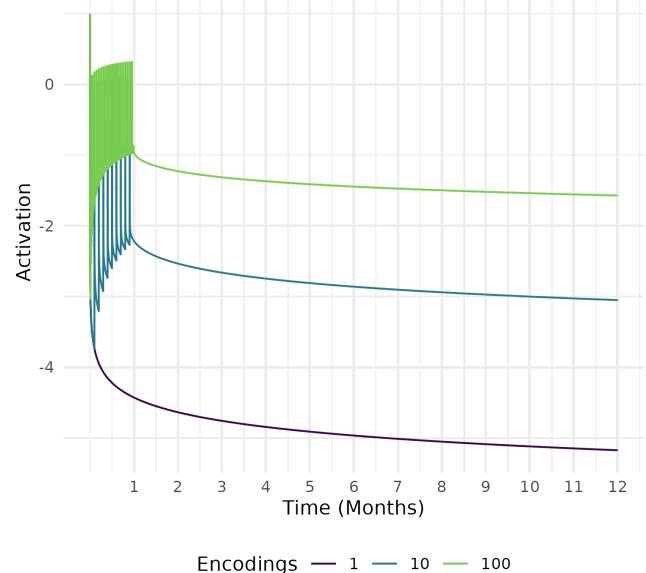Simulated Effect of Prior Knowledge Over Time

2

Figure 2: Simulated activation of three memories as a function of time following different numbers of encodings (1, 10, or 100) in the first month.

Materials and Methods

Participants

Undergraduates enrolled at the local university ($N = 70$, 46 Female, aged 18-21) were recruited on a rolling basis over a quarter to complete the study virtually. The recruitment criteria were as follows: (1) ages ranging from 18 to 29, (2) fluency in English, and (3) absence of significant medical or psychiatric conditions that could influence cognitive abilities. Participants who completed the prior knowledge survey and both fact-learning tasks were provided with course credit as compensation.

Prior Knowledge Survey

A PsychoPy task was designed to collect ground-truth data for participants' prior knowledge. Thirty national flags were used, pulled from the Caribbean Flags and Asian Flags learning lessons in a prior study. These flags provided an adequate range of potential prior knowledge, with flags that were more likely to be known beforehand, such as Japan, mixed with countries less likely to be known, like Montserrat.

Prior Knowledge Probe

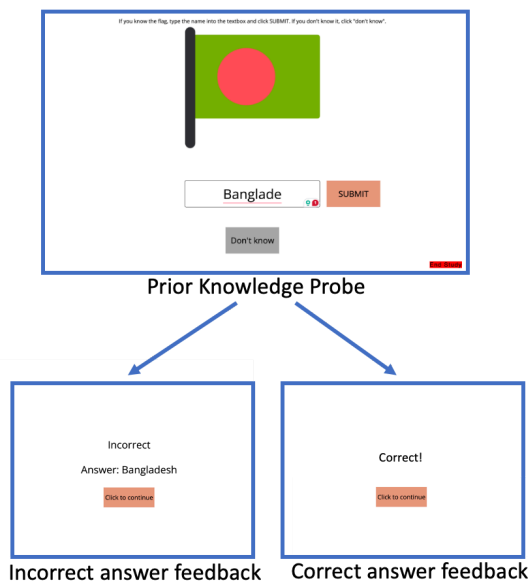Incorrect answer feedback   Correct answer feedback

Figure 3: Prior knowledge survey interface. Test screen followed by feedback screen that indicated typed answers as correct or incorrect.

As shown in Figure 3, participants were instructed to type the country name of the flag prompt if they were familiar with it. If they did not know it, participants selected the "Don't Know" button. The reaction time for country name guesses and "Don't Know" button presses was recorded. Facts that were typed correctly in the survey, with tolerance

for spelling, were marked as having "prior knowledge". Participants were provided feedback on the accuracy of their answers and provided with the correct answer in the case of an incorrect response. As such, this prior knowledge survey served a dual purpose to gauge ground truth data for the facts known beforehand and also to function effectively as the first trace of the fact.

Adaptive Fact-Learning System

Next, each participant completed two learning lessons, Asian Flags and Caribbean Flags, that were administered using a commercially available interactive interface (MemoryLab; Figure 4). This system dynamically estimates participants' *SoF* in real-time as they learn each stimulus-response pair (i.e., flag and country name). For each trial, participants answered a multiple-choice question and were shown the accuracy of their response. Each lesson was 8 minutes and consisted of 15 facts. Participants were not given a "study trial" as in prior studies due to the feedback given during the prior knowledge survey. It is important to note that the system's algorithm estimates parameters for the participants during the experiment, and these estimates are subsequently compared to the model including the prior knowledge constant. This assessment is described further in Sense et al. (2016) and can be accessed at https://www.memorylab.nl/en/.

Multiple-choice probe

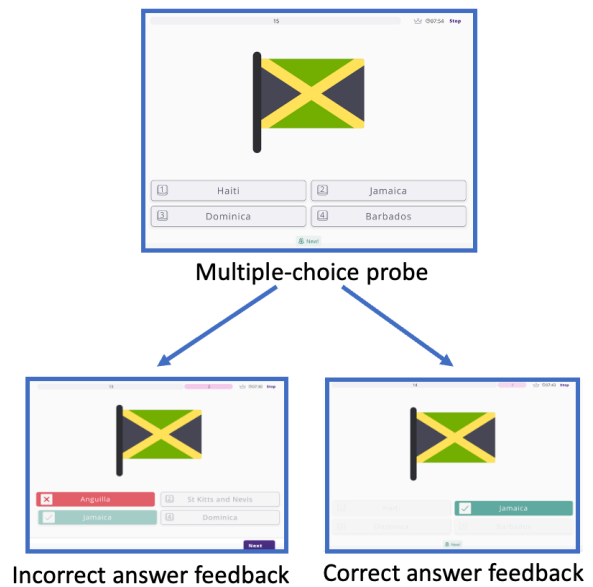Incorrect answer feedback   Correct answer feedback

Figure 4: Interface of MemoryLab adaptive fact-learning software. Presentation of multiple-choice questions, correct answer feedback, and incorrect answer feedback, shown respectively.

Maximum Likelihood Recovery of Prior Knowledge

To recover the amount of prior knowledge associated with each fact learned by an individual, Maximum Likelihood Estimation (MLE) was used. The model was retrospectively

3

fitted to each individual's data by choosing the set of parameters θ that maximized the model's likelihood. These parameters represent both individual participants' characteristics and the "unknown" pre-testing activation of each fact that was presented to the participant during the experiment. Formally, the likelihood of a set of parameters given a vector of data $\mathbf{x}$, $L(\theta|\mathbf{x})$, is the probability of observing the data $\mathbf{x}$, given the model: $L(\theta|\mathbf{x}) = P(\mathbf{x}|\theta)$. Because our data consists of multiple independent responses $x_1, x_2, \ldots, x_N$, the likelihood can be expressed as the product of the probabilities associated with each response:

$$L(M|\mathbf{x}) = P(x_1|M) \cdot P(x_2|M) \cdot \ldots \cdot P(x_N|M) = \prod_i P(x_i|M)$$

Because the product of probabilities becomes vanishingly small, it is common to use *log*-likelihood:

$$\log L(M|\mathbf{x}) = \log \prod_i P(x_i|M) = \sum_i \log P(x_i|M)$$

In our case, each model was simultaneously fitted to two behavioral measures for each response: accuracy and its corresponding response time. Trial-by-trial probabilities for responses and response times were calculated as follows: Given that a memory's activation reflects its log odds of retrieval, the probability $P_t(m)$ that a memory m would be retrieved at time t is given by:

$$P_t(m) = 1 / (1 + e^{-A(m, t) / s})$$

where s is a noise parameter that follows a logistic distribution with standard deviation of $\sqrt{3}/\pi s$. Thus, given the state of the model, it is possible to compute the probability associated with each response. The probability of a correct response is $P_t(m)$, and the probability of an incorrect response is $1 - P_t(m)$.

When considering response times, the calculations become more complicated. In ACT-R, the response time RT associated with the retrieval time of a memory m is an inverse exponential function of the memory's activation:

$$RT = T_{ER} + F \cdot e^{-A(m, t)}$$

where T_{ER} is the non-retrieval time (e.g., the time needed to for perceptual and motor responses) and F is an idiographic free parameter. Note that this expression is deterministic; to transform it into a probability distribution, we must consider the distribution of noise around the activation. As noted above, noise s follows a logistic distribution. Therefore, the resulting probability distribution for response times is a shifted log-logistic distribution with parameters $\alpha = e^{-A(m)}$ and $\beta = \sqrt{3}/\pi s$:

$$P(RT) = (\beta / F\alpha)((t - T_{ER})/\alpha)^{\beta-1} / (1 + (t - T_{ER})/(F\alpha)^\beta)^2$$

With these equations in place, it is possible to run a maximum likelihood estimation procedure to recover the most likely BLC values for every study item in a memory

experiment. The full model has one parameter (K) for each fact, and four parameters for each individual: φ, F, s, and T_{ER}. However, the adaptive fact learning system maintains F, s, and T_{ER} to constant defaults ($F = 1$, $s = 0.25$, and $T_{ER} = 300$ms). We will adhere to the same principle. Because no known closed-form formula exists to estimate the maximum likelihood solutions for this model, we used a derivative-free numerical minimization procedure, the simplex method (Nelder & Mead, 1965), as implemented in the Python SciPy package. To address the potential difference between the first "study trial" and the beginning of the learning trial, the offset time was calculated using computer timestamps and integrated to account for potential decay in activation.

Experimental Hypothesis

Based on the considerations laid out in the introduction, a number of experimental predictions can be made. First, we predict that prior knowledge of an item would be inaccurately estimated with a lower *Speed of Forgetting* (SoF). This arises from the assumption of a base-level activation (BLC), being equal to 0 for all facts, resulting in the model erroneously inferring quick and easy learning for previously known facts, thereby underestimating *SoF*.

We also predict that a weak correlation will exist between *SoF* values for previously known and previously unknown facts across participants. This correlation is expected due to the participant's *SoF* still influencing the benefit gained from multiple probes, even when a fact is already known.

Finally, we expect that a MLE-based parameter recovery procedure would be able to correctly identify previously known items as having large BLC values and that an automatic prior knowledge detector, utilizing a simple threshold model, could achieve greater than chance accuracy in identifying these facts.

Results

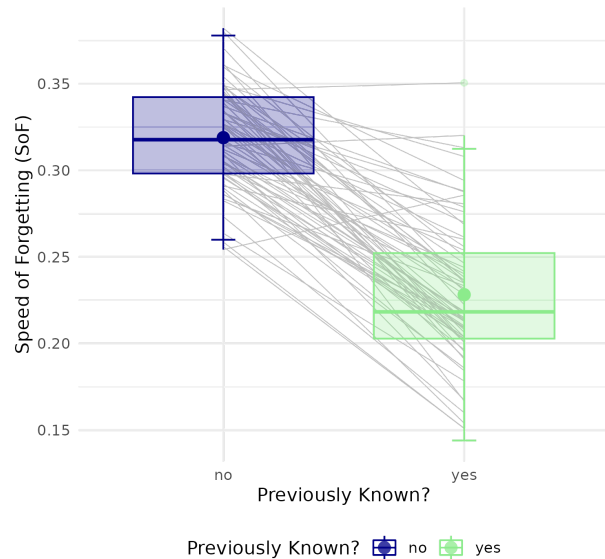Effect of Prior Knowledge on SoF Estimates

4

Figure 5: Effect of Prior Knowledge across participants. Facts known beforehand (green) had consistently lower *SoF* than facts not known beforehand (blue). Gray lines represent individual participants; colored boxes represent Tukey's boxplots, colored dots represent mean values.

Effect of Prior Knowledge on *Speed of Forgetting*

We found a significant impact of prior knowledge on the estimates of *Speed of Forgetting*. Notably, items identified by participants as previously known were consistently estimated to have a lower *SoF* compared to unknown items. This trend was evident across all participants, as illustrated in Figure 5. Utilizing a random slope linear mixed-effects model, we further analyzed this effect and discovered that items designated as previously known were associated with a significantly lower *SoF* (β = -0.09, 95% CI [-0.10, -0.09], $p < 0.001$). For context, participants had prior knowledge for approximately 18.6% of the facts they were tested on.
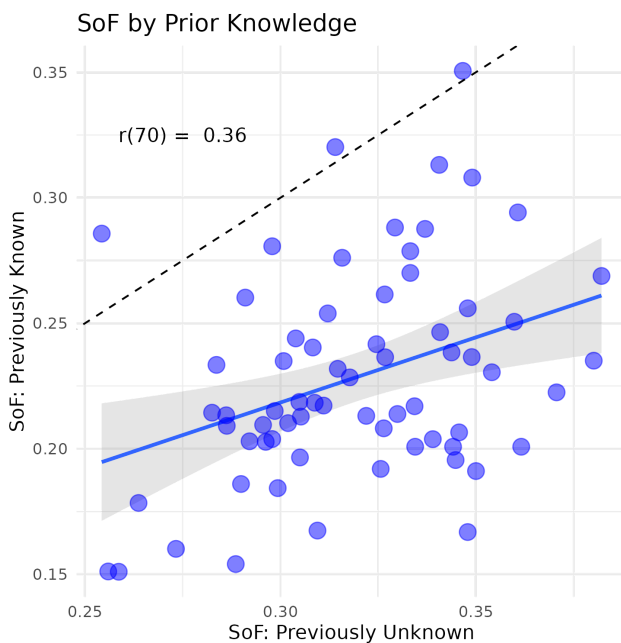

Figure 6: Correlation between the average *SoF* estimates for facts known beforehand vs. not known beforehand within each participant. Each point represents a participant in the study; the dashed line is the identity line.

Correlation Between Known and Unknown Facts

Our second hypothesis centered on investigating the relationship between *SoF* values for previously known and previously unknown facts across participants. We anticipated a weak correlation between these values, suggesting that the *SoF* of known facts would still influence subsequent assessments, even after initial familiarity. Indeed, our analysis revealed a small yet statistically significant correlation ($r(70) = .36$, $p = 0.002$; Figure 6), supporting our hypothesis and indicating the persistence of *SoF* effects even with prior knowledge.

Recovered Base-Level Constants

To test our third and final predictions, we conducted the MLE procedure on the dataset for each participant, recovering the most likely BLC value corresponding with each fact. While the model itself is blind to whether a fact was previously known or not, it correctly estimated that, on average, the BLC values for previously known facts were much higher than those for unknown facts (paired $T(69) = -17.00$, p < 0.0001) as shown in Figure 7. Importantly, the mean BLC values were higher for previously known facts across *all* participants, and were correctly estimated as close to zero for most previously unknown facts.
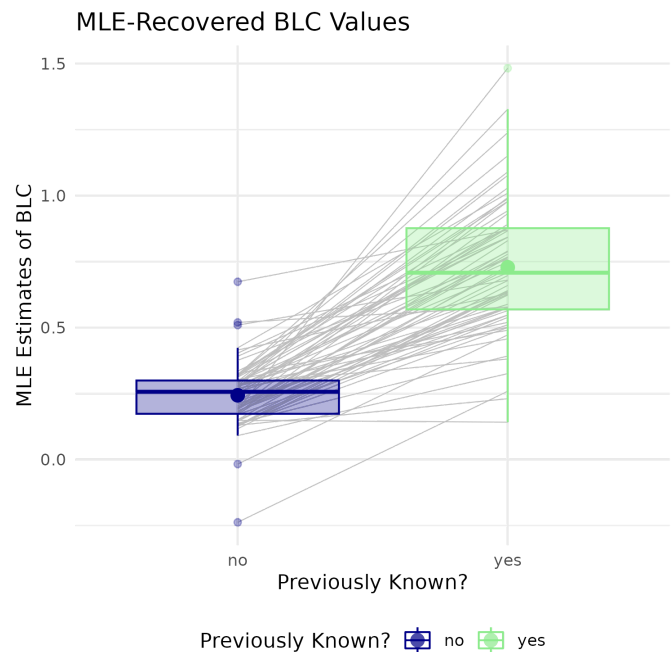

Figure 7: Recovered Base Level Activation values for facts with true prior knowledge (green) vs. without true prior knowledge (blue). Gray lines represent values for individual participants; colored lines represent Tukey's boxplots; colored points represent means.

Figure 7 shows that an automatic procedure can correctly identify previously known items, but it does not provide a quantitative measure of its efficacy. To obtain this estimate, we transformed the continuous estimates of BLC values into binary predictions by applying a moving threshold: items whose recovered BLC value was greater than the threshold were classified as "previously known". For each level of the threshold, it is possible to compare the classification accuracy against the ground truth, and estimate the proportion of true positives (hits) and true negatives (correct rejections) made.

These threshold-dependent proportions can be plotted to obtain the Receiver Operating Characteristic (ROC) curve, as shown in Figure 8. The area under the curve (AUC) of such a curve represents the classification accuracy of the method. In this case, the AUC is 0.81, implying that the

5

naive binary classification based on MLE-recovered BLC values alone could achieve 81% accuracy in identifying previously known items. Simply put, the *SoF* error from prior knowledge could be identified automatically and with above-chance accuracy *without* knowing which facts participants knew beforehand.
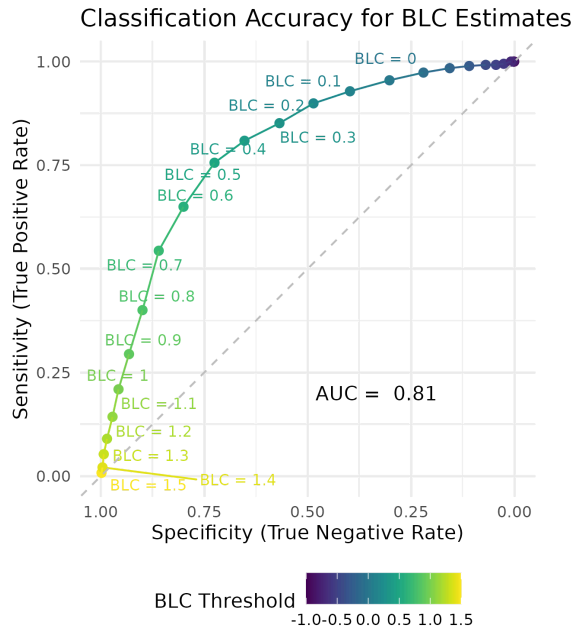

Figure 8: ROC-curve for inferring prior knowledge given the recovered base-level activation from MLE.

Discussion

The current study introduces a novel approach for disentangling prior fact knowledge in memory assessments from previously collected data. Our results confirmed the hypothesis predictions, demonstrating a significant decrease in the estimations of *Speed of Forgetting* (SoF) for items with prior knowledge. Furthermore, we observed a weak correlation between *SoF* values for facts with and without prior knowledge. Notably, we developed a simple threshold model that accurately predicted prior knowledge using Maximum Likelihood Estimation (MLE) and response times with 81% accuracy. This model-based approach effectively identified and dissociated the confounding influence of prior knowledge, representing a significant advancement in memory assessment methodology.

Identifying prior knowledge as a source of error in model-based assessments is a novel and important finding with implications across multiple levels of analysis. When it is unfeasible to account for prior knowledge initially, such as in previously collected data, it is possible to parse out individual facts that likely have prior knowledge. However, with 81% accuracy, this approach may not guarantee an improvement in model accuracy for the intention of classifying memory ability. The second level of analysis would be to include a model parameter K_m as discussed previously. Prior knowledge of a fact would alter base-level

activations accordingly. The third level would be to actively collect prior knowledge data. During the fact-learning task, participants would have the option to click "I already knew that" in study trials. When this happens, the base-level activation for that would be adjusted. With these model improvements, we predict that the memory classification accuracy will improve.

To address some limitations, the sampled age population likely has extensive experience taking multiple-choice examinations and could have developed alternative cognitive strategies for picking the correct answer other than successful retrieval, such as process of elimination. The multiple-choice format of the task was chosen due to previous use in the literature displaying more consistent *SoF* values (Sense et al., 2016). Finally, as memory is certainly related to age, the sample in this study does not reflect the elderly population that the task would be used for in clinical settings. A geriatric population could have more or less error resulting from prior knowledge, the results must be carefully considered before extrapolating to other age ranges.

Implications and Future Directions

Memory impairments are a common and debilitating aspect of aging, particularly in neurodegenerative conditions. The ability to quantify individual differences in memory is crucial, as early detection of memory impairment is essential for effective treatment. Moreover, the brief and user-friendly online format makes the administration of assessments remarkably convenient.

In light of our study's findings, it is imperative to minimize error in memory assessments. Our identification of error stemming from prior knowledge significantly affecting estimated *SoF*s highlights the potential for misclassification of individuals, particularly elderly participants with prior knowledge, as healthy controls rather than memory-impaired individuals. Addressing this source of error reduces the risk of Type II errors in memory assessment tasks, thereby enhancing the accuracy of diagnostic outcomes.

To further enhance the accuracy of our model-based assessment, we are undertaking another study aimed at better understanding the model parameter T_{ER} by integrating eye-tracking data. Analysis of scanpaths extracted from eye-tracking data will enable a comprehensive examination of the components contributing to this parameter, ultimately leading to improvements in the model's accuracy. With these combined improvements, we aim to rival the classification accuracy of clinical-standard assessments, thereby facilitating timely intervention and leading to improved patient outcomes.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 6(4), 451–474.

6

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science*, *2*(6), 396–408.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102.

Hake, H. S., Leonard, B., Ulibarri, S., Grabowski, T., Van Rijn, H., & Stocco, A. (2023). Breaking New Ground in Computational Psychiatry: Model-Based Characterization of Forgetting in Healthy Aging and Mild Cognitive Impairment. In *medRxiv* (p. 2023.05.13.23289941). https://doi.org/10.1101/2023.05.13.23289941

Nadel, L., Samsonovich, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus*, *10*(4), 352–368.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699.

Nelder, J. A., & Mead, R. (1965). A Simplex Method for Function Minimization. *Computer Journal*, *7*(4), 308–313.

Newell, A., & Rosenbloom, P. S. (1981). *Mechanisms of skill acquisition and the law of practice*. https://doi.org/10.4324/9780203728178-1/mechanisms-skill-acquisition-law-practice-allen-newell-paul-rosenbloom

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, *8*(1), 305–321.

Sense, F., & van Rijn, H. (2022). Optimizing fact-learning with a response-latency-based adaptive system. *Psyarxiv.com*. https://psyarxiv.com/chpgv/download?format=pdf

Somers, S., Oltramari, A., & Lebiere, C. (2020). *Cognitive twin: A cognitive approach to personalized assistants*. https://ceur-ws.org/Vol-2600/paper13.pdf

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219.

Zhou, P., Sense, F., van Rijn, H., & Stocco, A. (2021). Reflections of idiographic long-term memory characteristics in resting-state neuroimaging data. *Cognition*, *212*, 104660.

Modeling Instance-Based Rule Learning in an Adaptive Retrieval Practice Task

Thomas Wilschut (t.j.wilschut@rug.nl)[1]**, Florian Sense**[2]**, Myrthe Braam** [3]**, Hedderik van Rijn**[1]

[1]Department of Experimental Psychology, University of Groningen; [2]Infinite Tactics, LLC;
[2]MemoryLab BV

Abstract

Model-based adaptive learning systems have successfully improved the efficiency of fact learning in educational practice. Typically, such systems work by keeping track of a learner's memory processes by measuring behavior during learning, and using this information to tailor the learning process towards the needs of individual learners. Where many adaptive learning systems applied today focus on learning paired associates, we here focus on learning grammar rules based on instances of these general rules. We show that participants' (N = 42) behavioral responses on instance questions for a rule can be used to infer general performance on other questions associated to that rule, and that we can capture this rule performance in a single model-based *speed of forgetting* parameter. These findings could be used to develop and optimize adaptive learning systems that can be used to study general rules from instances.

Keywords: ACT-R; Adaptive Learning; Knowledge Tracing; Instance-Based Learning; Grammar

Introduction

Adaptive learning systems have successfully improved the process of memorizing factual information, such as vocabulary or glossary items, by tailoring learning schedules to the needs of individual learners. Typically, such systems aim to predict learner performance from behavioral measures that are recorded during learning, and use these predictions to tailor item repetition schedules towards the needs of individual learners (e.g., presenting fewer or easier items when predicted performance is low; and presenting more or more difficult items when predicted performance is high). This approach has proven to increase learning efficiency compared to traditional, less adaptive approaches in a wide range of materials, both in laboratory and classroom settings (e.g., see Lindsey, Shroyer, Pashler, & Mozer, 2014; Papousek, Pelánek, & Stanislav, 2014; Wozniak & Gorzelanczyk, 1994; Van Rijn, Van Maanen, & Van Woudenberg, 2009).

Existing adaptive learning systems are typically used to learn paired associates, such as vocabulary items or glossary items. For these materials, there is extensive evidence supporting the idea that it is possible to use behavioral proxies, recorded during the learning session, to infer the extent to which a learner has successfully memorised a specific paired-associate item. Most model-based adaptive learning systems present the learner with retrieval practice questions, and use response accuracy as a behavioral proxy of the extent to which an item is stored in memory (e.g., see Pavlik & Anderson, 2008; Van Rijn et al., 2009). As using accuracy scores only does not allow for meaningful discrimination within correct responses, (and as a consequence, accurate performance predictions require many incorrect responses,) many

systems use response times in addition to accuracy scores to predict performance (Byrne & Anderson, 1998; Sense, Behrens, Meijer, & van Rijn, 2016, see). Finally, in recent implementations, information carried in the speech signal during spoken retrieval attempts has been used to infer the extent to which a learner has successfully memorised a specific paired-associate item (Wilschut, Sense, & van Rijn, 2024). Overall, for learning paired-associate items, there is extensive support for the idea that behavioral responses to retrieval practice questions can be used to infer model parameters that map on to latent memory processes.

A popular framework used in model-based adaptive learning systems is the ACT-R model of human declarative memory (Anderson et al., 2004). In ACT-R, learners' memory representations for individual facts are stored as *chunks* in declarative memory. Chunks are schematic units of information that possess an activation value: More active chunks are more likely to be retrieved during a search of declarative memory. Arguably, a limitation of this model is that it treats individual facts as independent units of information. As such, it is not straightforward to model a learner's memory for facts that are clustered or related to other facts that have been encountered in the learning session (although accounts of *spreading activation*, in which activation spreads through a semantic network, could account for such context effects (e.g., see Anderson, 1983; Thomson, Bennati, & Lebiere, 2014)).

Although model-based adaptive learning systems have proven to be successful in improving the efficiency of learning paired associates, it is unclear to what extent these findings generalise to situations where facts are not independent from each other (i.e., where the clustering of items plays an important rule). In this research project, we aim to extend existing adaptive learning models that keep track of memory performance for simple paired associates by modeling a learner's mastery of general/underlying rules from instances of that rule (i.e. instance-based learning, see Lejarraga, Dutt, & Gonzalez, 2012). If adaptive learning models are able to keep track of a learner's mastery of a common rule based on responses to instance questions, this would widen the scope of such systems and their possible application in a wide range of educational settings. For example, current teaching methods for learning language grammar rules, mathematics, physics or chemistry all heavily rely on teaching students to pick up regularities or general rules from instances.

There have been several successful attempts at modeling the process of learning common rules or patterns from a set of instances. For example, Stevens et al., 2018 showed that

it is possible to model others' decisions from instances in a negotiation task. Instance-based rule learning models have also been made for learning the English past tense (Taatgen & Anderson, 2002), the German plural (Taatgen, 2001), as well as for other domains, such as the balanced-scale task (Van Rijn, Van Someren, & Van der Maas, 2003). Finally, within the context of ACT-R, studies have focused on using instance-based learning to explain human decision making (e.g., see Gonzalez, Dutt, & Lebiere, 2013). Yet, the above approaches all aim to model *inductive* rule learning from instances. In other words, the rule is never explicitly given to the learners. In the current work, we intend to explicitly provide feedback explaining the rule after each instance question, with the intention that the learners remember the rule, and recognize future instance questions that are associated to the same rule. To our knowledge, we are the first to model instance based rule learning in this exact setup.

In this project, we aim to explore if we can model instance-based rule learning in an adaptive retrieval practice task, where participants study Dutch grammar rules from specific instances. We specifically aim to track a learner's mastery of underlying rules, and therefore model these rules, and not the instances, as chunks in the memory model. We first examine the extent to which performance on instance questions for specific questions is associated to (a) other instance questions for the same grammar rule during learning and (b) new instance questions presented on a test following the learning session. Second, we will examine if we can use model-based estimates of speed of forgetting during learning to predict performance on the test. Finally, we aim to show that using a fully adaptive, model-based item scheduling algorithm—that takes both a learner's accuracy scores and response times into account to determine the most optimal item repetition schedule for each individual learner—can be used to successfully improve learning efficiency.

Methods

Participants

In total, 42 participants completed the experiment via the online participant pool *Prolific*. Participants were included if they had at least completed secondary education. Most participants had completed education at a university of applied science ('HBO'). In addition, they were required to speak Dutch fluently. Finally, participants were included only if they had completed at least 10 other Prolific studies prior to the current experiment. The mean age of the participants was 35 years, 18 participants identified as female and 24 participants as male.

Design and Procedure

The study consisted of two learning blocks and a test block, which were completed by all participants in a single session. All participants started with the learning blocks, which consisted of one rt-adaptive learning block and one stack-based learning block. Half of the participants (n = 21) started with the rt-adaptive learning block, and completed the stack-based block second. For the other half of the participants (n = 21), this order was reversed. After the learning blocks, a test followed.
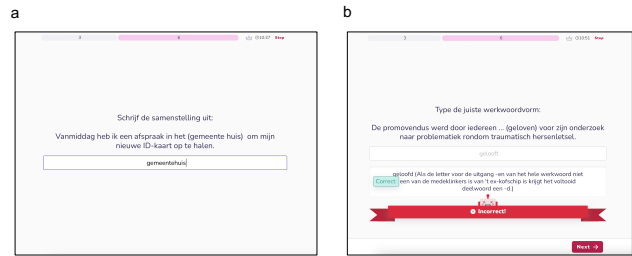
Figure 1: Screenshots of the learning application. **a** shows a trial: the learner is asked to type the correct word **b** shows the feedback when an incorrect answer is given: the correct answer is provided along with the explanation of the underlying grammar rule.

During the learning blocks, participants studied Dutch grammar rules based on instance questions (see Materials). For each grammar rule, there were six instances/instance questions, that were randomly chosen to be presented to the learner (with the only exception that a specific instance question would not be repeated twice in a row). Participants were prompted with a request (e.g., '*write the plural form of the word between brackets*') and a context sentence in which the target word occurred. Participants were asked to type the item in correct spelling. If the answer was correct, a short feedback screen appeared prompting the participants that the answer was correct, after which the next item was presented. If the answer was incorrect, the correct answer, as well as an explanation of the associated grammar rule, were presented to the participants. The feedback screen after incorrect responses was self-paced: Participants were able to click 'next' to continue at their own pace. Response times were defined as the time elapsed between the presentation of the cue and the first keypress, unless the participants used the backspace button to erase the first keypress (in which case the response time would not be used). Both the stack-based learning block and the rt-adaptive learning block took 12 minutes in total. In the rt-adaptive scheduling block, rule repetition schedules were personalised based on the accuracy and response times that were recorded during the learning session (see Rule scheduling); in the stack-based learning block, rules were scheduled based on accuracy only (see Rule scheduling).

On the test, one new instance question was presented to the participants for each grammar rule, where one rule was presented at the time. During the test, response times were not recorded.

Materials

The materials for this study were generated in collaboration with *Hogeschooltaal* (see https://www.hogeschooltaal.nl/?lang=en, a Dutch institution facilitating the process of language proficiency development in Dutch applied university students. The total material set consisted of 18 grammar rules, for each of which there were seven instance questions. Six instance questions were used in the learning session, one instance question was used for the test. All participants saw the same questions on the test. The list of 18 grammar rules was split in two sets of nine rules, which were then randomly assigned to a specific scheduling block (rt

adaptive or stack-based) for each participant.

Rule scheduling

In the rt-adaptive scheduling block, we used an adaptive algorithm to schedule rule repetitions in a way that is optimally tailored towards the individual learner. This adaptive algorithm is based on an ACT-R model of declarative memory (Anderson et al., 2004), and is described in more detail in Sense et al. (2016). In the current application, individual grammar rules—not individual instances—are stored as chunks in the declarative memory model. The algorithm aims to model the memory strength or activation of each to-be-learned grammar rule over time, and presents rules to the learner for retrieval practice whenever their activation decays to a threshold value. Activation values are continually updated using the learner's response times and accuracy scores.

In practice, this means that instances for which a learner gives slow and/or incorrect answers, activation values are adjusted downwards and rules are repeated more frequently, whereas if the learner gives quick and correct answers to a retrieval practice question, the activation will be adjusted upwards, and presented for practice less frequently. In addition to personalising the rule repetition schedule, the algorithm captures individual differences in ability through a learner- and item-specific *speed of forgetting* parameter (α), which it estimates from the learner's responses. Poorer learners will have a higher speed of forgetting value, which causes activation to decay faster, leading to more frequent repetition.

In the stack-based learning block, the rule repetition schedule was determined by a Leitner-inspired stack-based system (Mubarak & Smith, 2008), which groups words into virtual boxes: All words start in Box 1 and move to the next box if answered correctly. If a word is answered incorrectly, it moves back to the previous box. Words in Box 1 are presented first, followed by words in Box 2, followed by words in Box 3. If all rules are in Box 3 (and if they are all answered correctly) the rules are repeated in the order of first presentation until the learning time is over. This stack-based system allows for difficult rules to be rehearsed more often than easy rules and is a frequently used and effective study strategy (Bryson, 2012).

Analyses

Analyses were conducted in R 3.4.1 (R Core Team, 2020), with the mixed-effects modelling package lme4 1.1-28 (Bates, Mächler, Bolker, & Walker, 2015). The mixed effects models reported in this study include rule repetition, scheduling algorithm (contrast coded: rt-adaptive learning = 0; stack based learning = 1) and speed of forgetting. In all models, participant- and rule id were added as random intercepts (Baayen, Davidson, & Bates, 2008). The data was visualised using ggplot2 (Wickham, 2016).
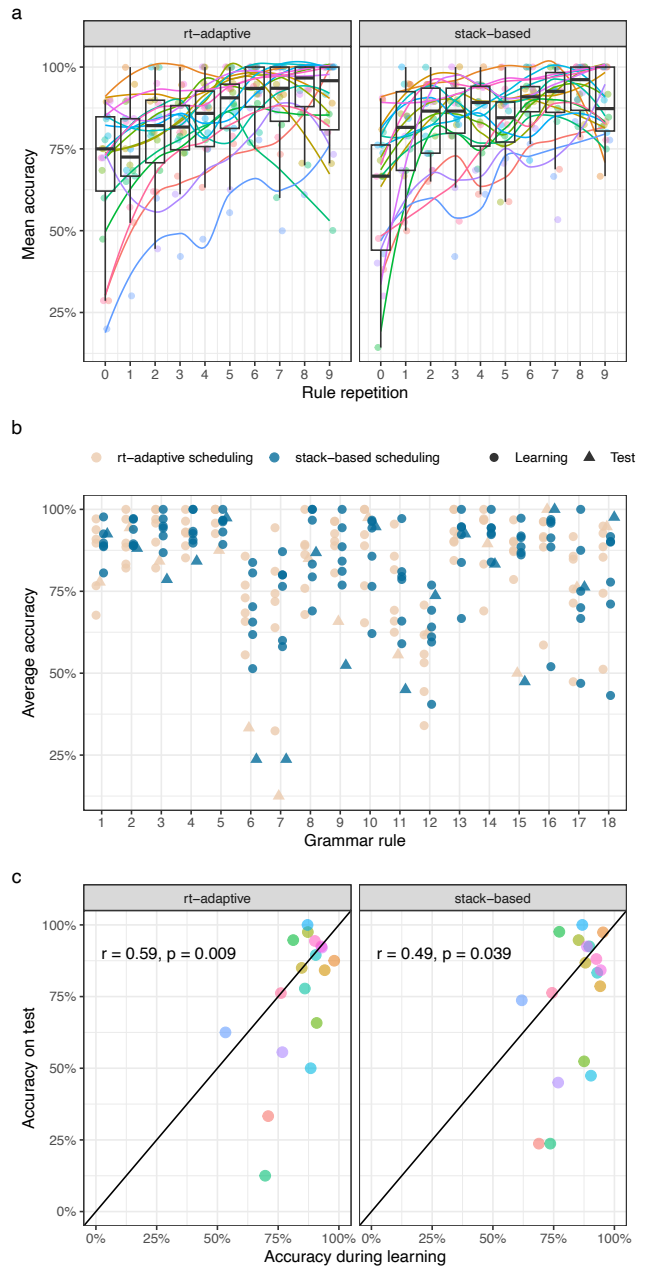

Figure 2: Inferring grammar rule mastery from instance-based learning behavior. Different colors represent unique grammar rules. **a** shows the mean accuracy over repetitions of a grammar rule, split by scheduling algorithm. Dots represent aggregate performance over randomly introduced instance questions for each rule. The graph shows that rule difficulty can be relatively reliably inferred from average scores on specific instances of that rule. For simplicity, only the first six of 18 grammar rules are shown here. **b** shows the main accuracy for each individual instance question during learning (dots) and associated test performance (triangles). **c** shows the association between mean accuracy during learning on instance questions associated to specific rules and accuracy during the test on new instance questions for the same rules.

Results[1]

Inferring grammar rule performance from instance-based learning behavior

The first aim of this research project was to examine the extent to which it is possible to use behavioral responses to instance questions to infer a learner's mastery of an underlying grammar rule. Figure 2a shows the mean accuracy on grammar rule questions over repetitions, split by scheduling algorithm. Colored lines represent individual grammar rules, and dots show average scores on repetitions of each rule, which are based on aggregates over instance questions. The figure clearly shows that it is possible to distinguish trends of rule difficulty: some grammar rules, aggregate accuracy scores over instance questions are lower than for other grammar rules. Correspondingly, for some grammar rules, initial performance is very low (close to 0), whereas for other rules, initial performance is quite high. Finally, the plot shows a trend of learning over repetitions (i.e., on average, accuracy increases over repetitions).

Figure 2b shows performance on individual instance questions for grammar rules. There is considerable variation between instance questions, but it seems reasonable to determine overall rule difficulty from a few observations of individual instances. Figure 2c shows that there is a strong positive association between average accuracy for grammar rules during learning and accuracy on new instance questions for the same grammar rules on the following test, both in the rt-adaptive scheduling block ($r = 0.59$, $p = 0.009$) and in the stack-based scheduling block ($r = 0.49$, $p = 0.039$).

Mixed effects models M1 and M2 (see Table 1 describe the effects of repetition, scheduling system, and their interaction on learning accuracy and response times, respectively. We found only significant main effects of repetition: participants became more accurate and responded faster over repetitions of a rule, regardless of the rule scheduling algorithm and despite the fact that rule repetitions consisted of randomly chosen instance questions. The effects of rule scheduling algorithm were not statistically significant. Overall, behavioral responses on instance questions for grammar rules seem to be indicative of performance on other instance questions that are associated to the same rule, both during the learning sessions and on the test that follows learning.

Model-based estimations of test performance

The second aim of this project was to capture rule mastery in a model-based *speed of forgetting* parameter. Figure 3a shows the mean estimated speed of forgetting over repetitions of grammar rules, based on instance questions for each rule. With each rule repetition, the estimated speed of forgetting was updated based on the accuracy and response time of the learner's answer (see Rule scheduling).
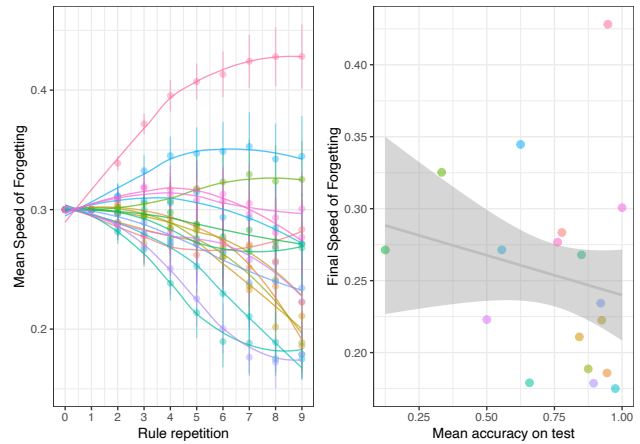
Figure 3: Estimating test performance based on model-inferred speed of forgetting for grammar rules. **a** shows the mean estimated speed of forgetting for each individual grammar rule, based on accuracy scores and response times for instance questions. Error bars represent (+/-) 1 standard error of the mean. **b** shows the mean test accuracy as a function of final speed of forgetting.
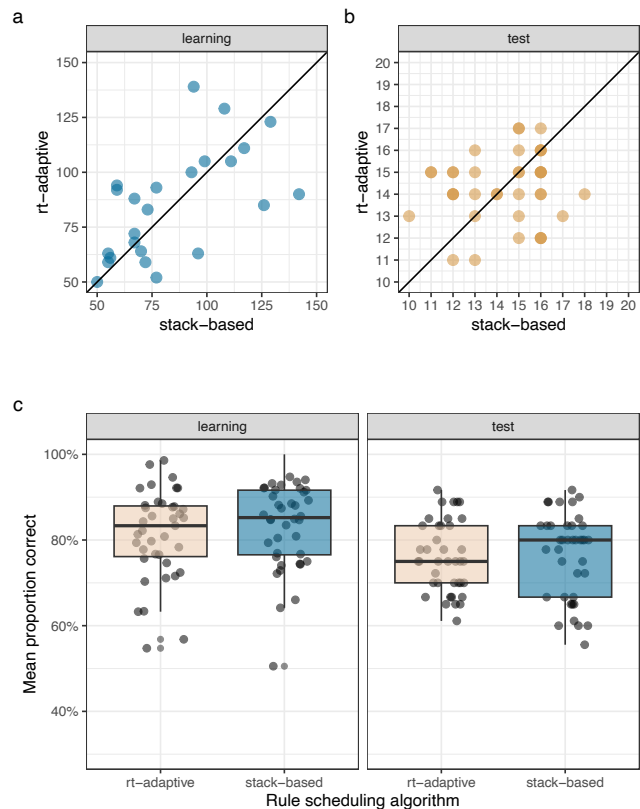

Figure 4: Performance during learning and test. **a** and **b** show the number of rules correctly recalled during learning and on test, respectively. Dots represent average scores for individual participants. **c** shows the proportion of correct responses during learning and on test for both rule scheduling algorithms.

Mixed effects models explaining performance during learning and on test from rule repetition and scheduling algorithm
*** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.05$.

M1. Accuracy during learning	Estimate	SE	z	p
Intercept	1.37	0.23	5.87	<0.001***
Rule repetition	0.08	0.01	8.85	<0.001***
Scheduling algorithm (stack-based = 1)	−0.06	0.01	−0.56	0.576
Rule repetition * Scheduling algorithm	0.018	0.01	1.28	0.212
M2. Reaction times during learning	Estimate	SE	t	p
Intercept	7711.91	1473.33	5.23	<0.001***
Rule repetition	−195.40	34.22	−5.71	<0.001***
Scheduling algorithm	−562.45	362.16	−1.55	0.121
Rule repetition * Scheduling algorithm	74.72	44.99	1.66	0.097
M3. Accuracy on test	Estimate	SE	z	p
Intercept	1.14	0.42	2.69	0.007
N. repetitions during learning	0.01	0.00	2.57	0.010*
Scheduling algorithm	−0.13	0.26	−0.50	0.616
Rule repetition * Scheduling algorithm	0.00	0.01	0.52	0.601

Estimating learning and test performance from rule repetitions and model-based speed of forgetting
*** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.05$.

M4. Accuracy during learning	Estimate	SE	z	p
Intercept	1.46	0.23	6.43	<0.001***
Speed of Forgetting	−0.84	0.21	−4.09	<0.001***
Rule Repetition	0.10	0.01	12.91	<0.001***
M5. Mean accuracy on test	Estimate	SE	t	p
Intercept	0.96	0.11	8.51	<0.001***
Final Speed of Forgetting	−0.76	0.31	−2.45	0.01*
N. repetitions during learning	0.00	0.00	0.71	0.47

Figure 3b shows the association between the final estimated speed of forgetting for a grammar rule during the learning session, and the mean accuracy during test for new instances of the same rules. We found that, overall, grammar rules for which a high speed of forgetting was estimated during learning, new instance questions were answered with lower accuracy on the test, indicating that the adaptive learning model could track rule performance during the learning session to estimate later test performance. Mixed effects models M4 and M5 (see Table 2) support these interpretations, as they show that the speed of forgetting for a grammar rule, estimated during learning based on responses to instance questions, can be used to estimate accuracy during learning, and on test, respectively.

Model-based optimization of learning

The final aim of this project was to explore the possibility of using the instance-based estimations of speed of forgetting for grammar rules to optimize repetition schedules, ultimately leading to a higher learning efficiency of grammar rules. To that end, we compared the learning efficiency with a fully adaptive scheduling algorithm, that uses both accuracy scores and response times to predict rule performance, to a stack-based rule scheduling

algorithm that is based on the accuracy of rule instances only (see Rule scheduling). Figure 4a and 4b show the number of correct responses during learning and on test, respectively. Figure 4c shows the proportion of correct answers during learning and test with both scheduling algorithms. As is also supported by the mixed effects models M1 and M3 (see Table 1), we found no significant difference between using the stack-based scheduling system and the rt-adaptive scheduling system.

Discussion

In this study, we aimed to extend existing adaptive learning models that can keep track of memory performance for simple paired associate stimuli to estimating mastery of grammar rules, based on responses to randomly introduced instance questions. The results can be summarized in three main points. First, we examined the possibility of inferring rule performance from behavioral responses on randomly chosen instance questions. Our results suggest that, despite the fact that we found considerable variation in performance on individual instance questions within a rule, it seems sensible to keep track of a learner's mastery of a grammar rule using the behavioral responses on randomly chosen instance questions. More specifically, accuracy scores on instance

questions for grammar rules were indicative of performance on other instance questions that are associated to the same rule, both during the learning sessions and on the test that follows learning.

Second, we showed that the adaptive learning model, that was originally developed to keep track of memory performance of individual paired associate rules, could capture the extent to which learners have mastered a grammar rule in a single speed of forgetting parameter. A higher speed of forgetting during learning for a specific grammar rule was associated with poorer memory performance for new instances of the same grammar rule.

Finally, we attempted to use the model-based estimations of a learner's performance of a certain rule by optimizing the repetition schedule for individual learners. We found that, despite the fact that our learning model could capture differences between grammar rules, model-based optimization of the rule repetition schedule did not lead to better learning performance compared to a more simple, stack-based adaptive learning system.

There are several possible reasons for the lack of a benefit of using the model-based rt-adaptive scheduling algorithm compared to using a stack-based accuracy adaptive system. First, in the current system, response times were defined as the time elapsed between the first presentation of an instance question and the first keypress of the response. The underlying rationale is that this response time mainly reflects *retrieval time*, and can therefore be used as a proxy of the memory strength for a specific rule (Byrne & Anderson, 1998). This way of measuring response latencies has proved to be effective for paired associate learning, but it is possible that response times should be decomposed more carefully when it comes to grammar rule learning. For instance, future research should examine whether the non-retrieval time (i.e., the time needed to process a question before retrieval takes place, or the time needed to prepare a response after retrieval has taken place) can be subtracted from the response times before being taken into account to determine scheduling for more complex materials such as grammar rule learning. Second, it is possible that the current experimental setup was not sensitive enough to statistically detect differences in learning efficiency between the accuracy-adaptive stack-based and the model-based, rt-adaptive scheduling system. Future studies should further examine this issue, in particular over multiple learning sessions and including longer-term retention tests.

Another possible direction for future studies is taking a data-driven approach of clustering items, rather than defining the common grammar rules beforehand. A post-hoc *k*-means clustering analysis of the current dataset suggests that only 5–7 clusters is enough to accurately describe the variability of performance on instance questions. In other words, learners performed very similar on some of the grammar rules, which makes the usefulness of treating them as separate knowledge chunks questionable. As in some situations it might be difficult to establish the most optimal common rule clustering upfront, it may be worthwhile exploring methods to use a data-driven approach to group items for individual learners in real time, and then track a learner's progress on each group of items.

Another important point that has received little attention in the current work concerns the explanatory feedback about the grammar rules that was shown to the learner after each incorrect response. Future work should examine the consequences of providing explanations of grammar rules after each response, and how the time taken to study these rules during the feedback moments impacts learning efficiency.

Despite these open questions, we show that it is possible to model learners' mastery general rules from answers to instance questions, and that we can use this information to optimize rule repetition schedules. These results demonstrate that—in the context of learning Dutch grammar rules—it is sensible to use performance on instance questions to infer a learner's mastery of the underlying rule. Despite the fact that our current attempts at using this information to personalise the repetition schedule did not result in increased learning efficiency, our results indicate that it is sensible to track rule performance from responses on corresponding instance questions. These findings underline the need to further investigate possible ways of using this information to improve repetition schedules for these rules. Ultimately, this could lead to learning systems that allow for instance-based rule learning, adapted to the needs and prior knowledge of individual learners.

Conclusion

In this project, we asked participants to study Dutch grammar and spelling rules through exposure to specific instances of each rule. We show that it is possible to use the learner's answers to instance questions to estimate their performance on new instances of the same rules. Using a cognitive model of memory retrieval, we show that we can estimate how well learners have memorized the rules. Although future research should explore how these estimations of a learner's rule performance can be exploited to increase learning efficiency, these results pave the way for the development of adaptive learning applications that allow for rule learning based on instances.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, *22*(3), 261–295.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390–412.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bryson, D. (2012). Using flashcards to support your learning. *Journal of visual communication in medicine*, *35*(1), 25–29.

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. *The atomic components of thought*, *16*, 23–28.

Gonzalez, C., Dutt, V., & Lebiere, C. (2013). Validating instance-based learning mechanisms outside of act-r. *Journal of Computational Science*, *4*(4), 262–268.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological science*, *25*(3), 639–647.

Mubarak, R., & Smith, D. C. (2008). Spacing effect and mnemonic strategies: A theory-based approach to e-learning. In *e-learning* (pp. 269–272).

Papousek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. In *Educational data mining 2014* (pp. 6–13).

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in cognitive science*, *8*(1), 305–321.

Stevens, C. A., Daamen, J., Gaudrain, E., Renkema, T., Top, J. D., Cnossen, F., & Taatgen, N. A. (2018). Using cognitive agents to train negotiation skills. *Frontiers in psychology*, *9*, 154.

Taatgen, N. A. (2001). Extending the past-tense debate: A model of the german plural. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 23).

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? a model of learning the past tense without feedback. *Cognition*, *86*(2), 123–155.

Thomson, R., Bennati, S., & Lebiere, C. (2014). Extending the influence of contextual information in act-r using buffer decay. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Van Rijn, H., Van Maanen, L., & Van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th international conference of cognitive modeling* (Vol. 2, pp. 7–6).

Van Rijn, H., Van Someren, M., & Van der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. springer.

Wilschut, T., Sense, F., & van Rijn, H. (2024). Speaking to remember: model-based adaptive vocabulary learning using automatic speech recognition. *Computer Speech & Language*, *84*, 101578.

Wozniak, P. A., & Gorzelanczyk, E. J. (1994). Optimization of repetition spacing in the practice of learning. *Acta neurobiologiae experimentalis*, *54*, 59–59.

Challenges for a Computational Explanation of Flexible Linguistic Inference

Marieke Woensdregt (marieke.woensdregt@ru.nl)
Language and Computation in Neural Systems, Max Planck Institute for Psycholinguistics, The Netherlands
Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Mark Blokpoel (m.blokpoel@donders.ru.nl) **Iris van Rooij (i.vanrooij@donders.ru.nl)**
Department of Cognitive Science and Artificial Intelligence, Radboud University, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Andrea E. Martin (andrea.martin@mpi.nl)
Language and Computation in Neural Systems, Max Planck Institute for Psycholinguistics, The Netherlands
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, The Netherlands

Abstract

We identify theoretical challenges for developing a computational explanation of flexible linguistic inference. Specifically, the human ability to interpret a novel expression (like *mask-shaming*), where inferring plausible meanings requires integrating relevant background knowledge (e.g., COVID-19 pandemic). We lay out (i) the core properties of the phenomenon that together make up our construal of the explanandum, (ii) explanatory desiderata to help make sure a theory explains the explanandum, and (iii) cognitive constraints to ensure a theory can plausibly be realised by human cognition and the brain. By doing so, we reveal the 'force field' that theories of this explanandum have to navigate, and we give examples of tensions that arise between different elements of this force field. This is an important step in theory-development because it allows researchers who aim to solve one part of the puzzle of flexible linguistic inference to keep in clear view the other parts.

Keywords: language comprehension; inference; theory-development; computational explanation; meta-theory

Introduction

Language use is remarkably flexible. One aspect of this is that humans appear to be able to integrate different kinds of knowledge in novel ways when interpreting utterances. In this paper, we focus specifically on humans' ability to come up with possible interpretations of neologisms, such as *mask-shaming*. Coming up with a plausible interpretation of such a novel expression arguably requires an ability to relate knowledge of the meaning of the words and how they are combined, to broader contextual or world knowledge (e.g., about the COVID-19 pandemic; see Figure 1). Explaining this phenomenon raises several theoretical challenges[1]: What *is* the phenomenon really? What counts as a *good* explanation? The aim of this paper is to outline those challenges. Importantly, we consider how these challenges will interact, which brings into clear view a 'force field' that explanations of flexible linguistic inference need to navigate.

---

[1] We would like to preempt the possible presupposition that large language models (LLMs) would already address these challenges. LLMs do not provide any precise characterisation of the explanandum (human flexible linguistic inference), nor are they explanatory (Guest & Martin, 2023; Bender & Koller, 2020; van Rooij et al., 2023; van Rooij, 2022).

The contribution we make in this paper takes inspiration from several sources. First, Adolfi, van de Braak, and Woensdregt (2023) argue that theoretical problem-finding (as opposed to empirical problem-solving) is an important scientific contribution in its own right. This activity involves not just characterising the phenomenon, but also identifying the theoretical constraints that determine what makes a good explanation. Second, Guest (2024) and Guest and Martin (2023) argue that as scientific practitioners, we can make meta-theoretical commitments about criteria that make a theory good. Guest (2024) calls upon scientists to characterize and examine the criteria we use to adjudicate over theories by building and sharing what Guest and Martin (2023) and Guest (2024) dubbed a *metatheoretical calculus*: a formal system that describes the process by which theories are evaluated and pitted against each other in a particular (sub)field. Finally, Blokpoel (2018) argues that developing a computational-level model (i.e., a formalised theory) of a cognitive capacity is like sculpting. The scientist has to start out with a sufficiently large block of material (i.e., model/theory) that can capture the entire capacity (i.e., is *generatively sufficient*), and can then figure out which parts to chisel away by applying various *computational-level constraints* (e.g., tractability).

In this paper, we take inspiration from these approaches, and apply them specifically to the phenomenon of flexible linguistic inference. That is, the human ability to flexibly interpret neologisms upon first encounter, in a way that appears to require integrating linguistic knowledge with world knowledge. We start by outlining the specific phenomenon in language comprehension that we want to explain, in the form of three key properties, in Section *The explanandum*. Next, inspired by Guest (2024), Guest and Martin (2023), Blokpoel (2018), and Adolfi, van de Braak, and Woensdregt (2023), we discuss two classes of constraints (*Constraints on the explanans)* that we deem particularly relevant for theories of this explanandum. First, in Section *Explanatory desiderata*, we discuss two metatheoretical commitments that can help make sure a given theory really explains the explanandum of interest. Second, in Section *Cognitive constraints*, we discuss two metatheoretical commitments that can help make sure the
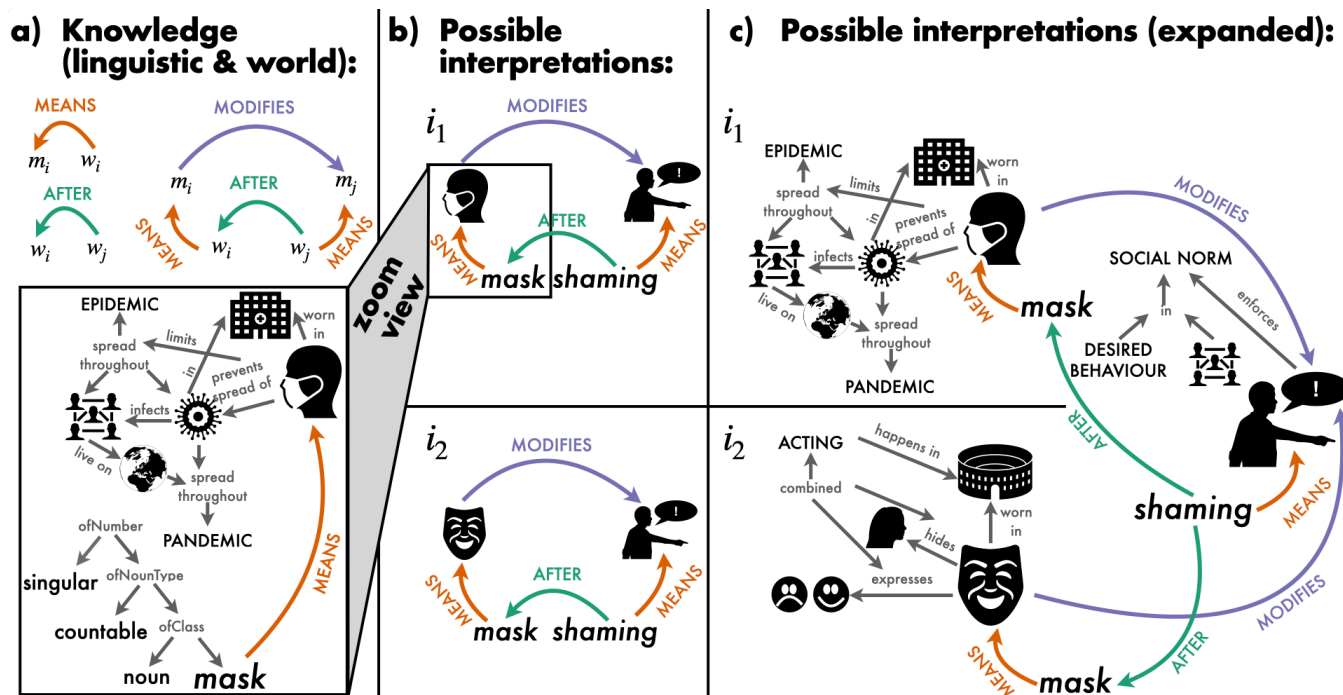
Figure 1: Illustrative example of our construal of the explanandum: The ability to come up with a plausible interpretation of the neologism *mask-shaming*. From left to right: (a) structured representations of stored knowledge (including grammatical, semantic, and world knowledge) are involved in building (b) structured representations of possible interpretations (here, two possible interpretations, i_1 and i_2 are shown, with a zoomed-in view of i_1 to illustrate the further structured knowledge that is associated with these abstract representations). Finally, (c) given the background knowledge associated with the semantic representations, and assuming this word is interpreted within the context of the COVID-19 pandemic, i_1 is more plausible.

theory can also be plausibly realised by human cognition and the brain. Finally, in Section *Challenges for explaining flexible linguistic inference*, we highlight some examples of tensions that may arise between these properties and constraints.

The properties of the phenomenon and constraints on the explanans that we highlight in this paper are not exhaustive; we see them as necessary but possibly not sufficient. However, by outlining the explanandum, several constraints on the explanans, and some of the tensions that can arise between these, we shed light on the force field that theories of flexible linguistic inference need to navigate. This provides a foundation from which further theory-development can depart.

The explanandum

In this section, we describe the core properties of the phenomenon we want to explain; that is: our construal of the *explanandum*. It appears to be the case that humans, under the right circumstances (given a shared language, shared world knowledge, and shared motivation to achieve mutual understanding) are able to interpret novel expressions in a way that requires knowledge not just of the word meanings and grammar of the language, but also broader contextual or world knowledge. And that these different kinds of knowledge are flexibly integrated in this process of meaning inference. For example, the first time you heard the term *mask-*

shaming, you were probably able to come up with a sensible (not necessarily correct) interpretation of what this might mean, in the context of the COVID-19 pandemic (Blokpoel, Wareham, Haselager, Toni, & van Rooij, 2019). Figure 1 shows a possible construal of what such an inference process might involve. Below, we discuss three properties that we believe together form the core of this explanandum. Our construal leaves out other components of language comprehension that also require explanation, such as segmentation (i.e., turning a continuous stream of sound or sign into discrete units; Adolfi, Wareham, & van Rooij, 2023) and word recognition (i.e., mapping a sequence of phonemes onto a lexical representation; Lahiri & Marslen-Wilson, 1991; McQueen, 2007). These capacities are outside the scope of this paper, as our construal of the explanandum does not rely on any particular theory of them.

Language comprehension is compositional

To understand the meaning of a linguistic expression (a phrase or sentence), one doesn't need to have come across it as a whole before. Instead, we can most often infer the meaning of the whole by knowing the meanings of the parts (lexical semantics) and how the structure of the whole influences its meaning (syntax). The fact that (most often) the meaning of the whole is a function of the meanings of the parts

and the way in which those are combined, makes natural languages *compositional* (Martin & Baggio, 2020; Partee, 1995; Pylkkänen, 2020). This compositionality buys us a high degree of systematicity and productivity (i.e., we can produce and understand utterances we have never come across before) (Szabó, 2004; Martin & Baggio, 2020; Pylkkänen, 2020).

Language comprehension requires building abstract hierarchical structure from linearly incoming sensory input, on the fly (Hagoort, 2019). Martin (2016, 2020) captures this computationally as a process of perceptual inference, in which incoming sensory cues are transformed into increasingly abstract structures through activation of stored knowledge representations. This computational model can account for cases of language comprehension in which the compositional meaning can be inferred directly from the stored language knowledge and its mapping to conceptual knowledge. However, humans are also able to infer the possible meanings of novel expressions in a way where semantics, syntax and compositionality alone are not enough.

Language comprehension involves world knowledge

Knowledge of the meanings of words (lexical semantics) is often not independent from world knowledge.[2] Hagoort et al. (2004) showed that in language comprehension, general world knowledge is integrated simultaneously with lexico-semantic knowledge (see also Hagoort & van Berkum, 2007). Using EEG, they showed that the event-related potential (ERP) component associated with semantic integration (the N400) looks similar in terms of timing, shape, and location when reading sentences like "the Dutch trains are white and very crowded" (a violation of world knowledge for the Dutch participants, who know that Dutch trains are yellow) compared to "the Dutch trains are sour and very crowded" (a semantic violation, because the semantic features of the predicate "sour" do not fit those of its argument "trains"). This is empirical evidence against the classic two-step model of language interpretation in which first the 'local' meaning of the compound expression is computed, and world knowledge is only integrated in a second step, to work out what the expression really means. Instead, Hagoort and van Berkum (2007) show that world knowledge is brought to bear on utterance interpretation as soon as it's available (Just & Carpenter, 1980; Hagoort & van Berkum, 2007; Hagoort, 2019).[3]

The importance of world knowledge for language comprehension becomes especially apparent when interpreting novel expressions such as *mask-shaming* (see Figure 1). We posit that in addition to building compositional structure based on stored and structured language and world knowledge, this re-

quires inferring new relationships between the incoming sensory cues and (potentially novel) conceptual representations. This may involve *abductive inference*, where novel candidate hypotheses to explain a given observation are generated (in this case: possible interpretations of a novel linguistic expression) (Blokpoel et al., 2019). Explaining this ability may require a computational model that can reach across different capacities in cognition and capture systematicity between structured representations of incoming language input and structured representations of world knowledge.

Language comprehension is incremental

Words (or signs) come in incrementally during language comprehension, in linear order (although signed languages allow for more simultaneity than spoken languages; Slonimska, Özyürek, & Capirci, 2020). Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) showed that linguistic utterances are also *processed* incrementally, not just syntactically but also semantically and in context. They showed that participants seek to establish reference in context immediately, as soon as words come in. More recently, Hagoort (2019) reviewed various psycholinguistic studies on meaning-making, and concluded that complex meaning is created on the fly, through a unification operation that takes lexical meanings and context as its input and outputs a situation model (see Zwaan & Radvansky, 1998, for more on situation models).

Sedivy (2007) reviews the psycholinguistic literature on incremental language processing in the context of theories of pragmatic inference (i.e., going beyond the literal meaning of an utterance to figure out what it means in context). She shows that there is at least some evidence from self-paced reading experiments that participants rapidly integrate expectations based on the informativeness of different possible referring expressions given the context. This suggests that also the pragmatic integration of context happens incrementally.

The combination of incremental and immediate language processing means that as hierarchical representations are being built from a linear input sequence, the set of possible interpretations and their hierarchical structure may change and need to be revised as new words come in. This means that a computational-level theory of flexible linguistic inference needs to be able to produce intermediate output when provided with only partial input sequences.

Constraints on the explanans

In this section, we highlight two classes of metatheoretical constraints or commitments that we deem particularly important for the explanandum described above. These two classes of constraints are somewhat different in nature: The *Explanatory desiderata* have to do with whether a given theory can really explain the phenomenon, and the *Cognitive constraints* have to do with whether the theory can be plausibly realised by human cognition and the brain. Blokpoel (2018) argues that developing a computational-level model (i.e., a formalised theory) of a cognitive capacity is like stone

---

[2]To illustrate how word meanings are underdetermined in the absence of world knowledge, Hagoort, Hald, Bastiaansen, and Petersson (2004) provide the following example: The word "finish" means something different in the phrase "Mary finished the book" (which implies she completed reading or writing it) compared to "the goat finished the book" (which implies the goat ate or destroyed it).

[3]For a computational model of this integration of world knowledge during incremental comprehension, see Venhuizen, Crocker, and Brouwer (2019).

carving a sculpture out of a block of marble. First, the modeller needs to make sure the block of marble they start with (the 'starting theory') is large enough to capture the entire explanandum (i.e., generatively sufficient). Otherwise, they would have to glue parts back on later, which, in this analogy, corresponds to adding ad-hoc components to the model. Second, they can start chiseling down the sculpture based on various constraints, until the model provides a precise fit of the cognitive capacity (i.e., the explanandum). In this paper, we build on this metaphor: We view the explanatory desiderata described below as characteristics of the block of stone that the sculptor starts out with, and the cognitive constraints as informing the chiseling process. This sculpting analogy also allows us to illustrate what consequences it has for the later chiseling process if the explanatory desiderata are violated.

Explanatory desiderata

Below, we discuss two metatheoretical desiderata that we consider important for a theory of a cognitive process to be explanatory. We also discuss the consequences if these desiderata are not satisfied: such a theory is likely to break apart during further chiseling based on cognitive constraints (Figure 2). This highlights the importance of having 'good quality material' to start with: A theory that (i) does not assume what it's trying to explain, and (ii) is not piecemeal.

Explaining without assuming Explanations of cognitive processes can be described on the computational level as a function that maps from input to output. That is, we can formalise a hypothesis about *what* a given cognitive capacity does (i.e., a computational-level explanation), as a function $f : I_f \to O_f$ that specifies for each input $i \in I_f$ its corresponding output $o \in O_f$ (Marr, 1982). Such a computational-level theory constrains the set of possible algorithmic-level and implementation-level specifications that are consistent with it (Blokpoel, 2018). By explaining without assuming, we mean that on the computational level, the theoretician should not slip by assuming that that which is to be explained is part of the input. Instead, the theory has to explain how a given property of the explanandum is part of the output *as a function of* the input. If, instead, this property that is in need of explanation is assumed, without explaining how it arises or where it comes from, the theory can be considered 'hollow', and this may reveal itself upon later chiseling.

Let us take the compositional nature of language comprehension as an example. The input in this case should be a linear sequence of words, and the output should be a hierarchical representation of the compositional structure that arises from the interaction between the meanings of the words and the way in which they are combined. If compositional structure is already present in the input to this function, it is assumed, rather than explained. If, instead, the formalisation of the model provides a specification of the output *as a function of* the input, where (some of) the output has compositional structure but the input does not, we can state that it explains compositionality without assuming it. Note that this definition of

explaining without assuming is independent of the specific definition of compositionality one is working with.

Non-piecemeal We consider a theory piecemeal if it makes use of different components (e.g., several separate computational processes) to explain different aspects of the explanandum of interest. The worry with such a piecemeal explanation is that it also requires an explanation of how these different components (e.g., computational processes) interface. This process of 'glueing' the different components back together may turn out hard (especially if these different component explanations were developed independently from one another), for example because they have incompatible assumptions. There can be valid reasons to conclude that a piecemeal explanation, postulating several different computational processes, is in fact necessary. However, aiming for a non-piecemeal approach first, can potentially avoid the hard problem of having to glue parts back together later. Furthermore, by adopting such a non-piecemeal approach, the limits of reaching such a unified, non-piecemeal explanation for a given explanandum will eventually be discovered, if they exist. This does require starting out with a well-specified and clearly carved out explanandum.

Let us take the different levels of organisation we find in linguistic expressions (from phonemes to morphemes to words to phrases to sentences) as an example. If our explanation entails a computational process that could be applied iteratively to build up interpretations from the smallest meaningful linguistic unit (morphemes) up to entire sentences, it can be considered non-piecemeal. If, instead, it has to postulate multiple computational processes in order to account for different levels of linguistic analysis, it is more piecemeal in nature. See Martin (2020) for an example of a non-piecemeal approach to explaining language comprehension.
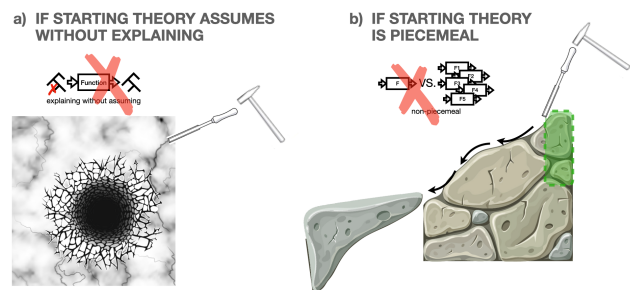

Figure 2: Illustration of how violating the two explanatory desiderata will affect theory-development during later chiseling based on cognitive constraints. a) When the theory *assumes* the explanandum, instead of explaining it, it can be seen as hollow, which may be revealed during the chiseling process. b) When the theory is piecemeal—made up of several components to explain different aspects of the phenomenon, without forming a coherent whole—it may break apart upon further chiseling (green outline indicates target area to be chiseled off). Stone images taken from freepik.com

Cognitive constraints

Here, we discuss two meta-theoretical constraints that are specific to theories that aim to explain cognitive capacities. These two constraints are necessary (i.e., any explanation that doesn't satisfy these two constraints will inherently not explain the phenomenon), but not sufficient (i.e., other constraints that we do not discuss may also apply, meaning any explanation that does satisfy the two constraints will not automatically provide a good explanation of the phenomenon). For example, given the explanandum of flexible linguistic inference, one may further want to ensure that the theory is consistent with insights from psycholinguistic research.

Computational tractability Human minds are resource-bounded. That is, we have limited time and memory resources. This means that human minds (just like any other resource-bounded system, such as a computer) can only compute *tractable* functions for real-world input sizes (as opposed to toy scale). This means that for an explanation of a cognitive capacity to be plausible, it has to be computationally tractable (van Rooij, 2008; van Rooij, Blokpoel, Kwisthout, & Wareham, 2019). One can analyse whether a given theory is tractable by specifying it at Marr's computational level (as a function that maps from input to output) (Marr, 1982), and using mathematical proof techniques from computational complexity theory (van Rooij, 2008; van Rooij et al., 2019).

If the theory of interest turns out to not be tractable, similar techniques from *parameterized* complexity theory can be used to find out whether the input domain can be constrained in a way that *would* make it tractable (van Rooij, 2008; van Rooij et al., 2019). By input domain here we mean anything that is part of the input to the function that describes the *what* of the cognitive capacity. Note that this is a different notion of *input* than the sensory input to the neural or cognitive system when we process language (e.g., the auditory input of a speech stream, the visual input of a sign or gesture stream, etc.). Instead, the input domain in this context also includes any stored knowledge that is used in the explanation of how the cognitive system gets to a certain output (e.g., an interpretation of a novel expression), such as lexico-semantic knowledge, grammatical knowledge, world knowledge, etc. For an example of such a parameterized tractability analysis applied to a theory of intentional communication that involves inferring others' communicative goals, see van Rooij et al. (2011).

Neural plausibility Computational tractability is analysed at Marr's computational level of analysis (Marr, 1982), and thus only requires a computational-level model that describes the *what* of the cognitive capacity in question. That is, describing the nature of the input-output mapping being computed (the cognitive *function*). However, as Martin (2016) argues, any model of language computation must not only answer such *what* questions, but also *how* questions. That is, to provide a specification at the algorithmic level of analysis: describing the nature of the algorithmic process by which the cognitive function is being performed (the cognitive *pro-*

cess). Similarly, Hagoort (2019) argues that the computational, algorithmic, and implementational level are interdependent, and that this should be taken into account when developing a mechanistic account of meaning-making in the mind (or in fact any cognitive function).

The set of possible algorithms is constrained by the computational-level explanation, but is also underdetermined by it (Blokpoel, 2018). That is, a given cognitive function (input-output mapping) can in principle be computed by different algorithms (van Rooij et al., 2019; van Rooij & Blokpoel, 2020; Blokpoel, 2018). However, as Martin (2016, 2020) demonstrates, algorithmic-level explanations can be constrained and informed by what we know about how the brain works: What type of computations can neural systems carry out? (e.g., summation and normalization.) (see also Martin, 2020; Kaushik & Martin, 2022). In addition to constraining possible theories to only those cognitive functions for which there exists an algorithm that can tractably compute it (see Section *Computational tractability*), one can further constrain the space of possible theories by putting additional constraints on the type of algorithm. Given a particular set of operators (e.g., summation and normalization) that are considered plausible for the brain to implement, one could make the commitment that the function needs to be computable by an algorithm that uses only these operators. In other words, one can make assumptions about the kind of architecture that cognition is implemented in, and make the commitment that the cognitive function a theory posits should be computable by an architecture of this type (van Rooij, 2008).

Challenges for explaining flexible linguistic inference

The sections above outlined the phenomenon to be explained, as well as the form that a good explanation should take, all together summarised in Figure 3. In the process of theory-development, tensions may arise between each of these properties and constraints. Figure 3 can thus be seen as describing a 'force field', within which tensions may arise both within and across levels. Below, we work out two of these tensions in a bit more detail: (i) explaining compositionality without assuming it (tension between a property of the phenomenon and an explanatory desideratum), and (ii) explaining the role of world knowledge tractably (tension between a property of the phenomenon and a cognitive constraint).

Explaining compositionality without assuming it

Explaining the compositional nature of language comprehension without assuming it raises questions for what type of linguistic knowledge can be considered part of the input domain (see Section *Computational tractability* for what we mean by input domain). There is a tension between assuming that the relevant grammatical knowledge is in place (i.e., that we're explaining flexible linguistic inference in competent adult language users), and providing an explanation of the computational operations that are necessary to get from
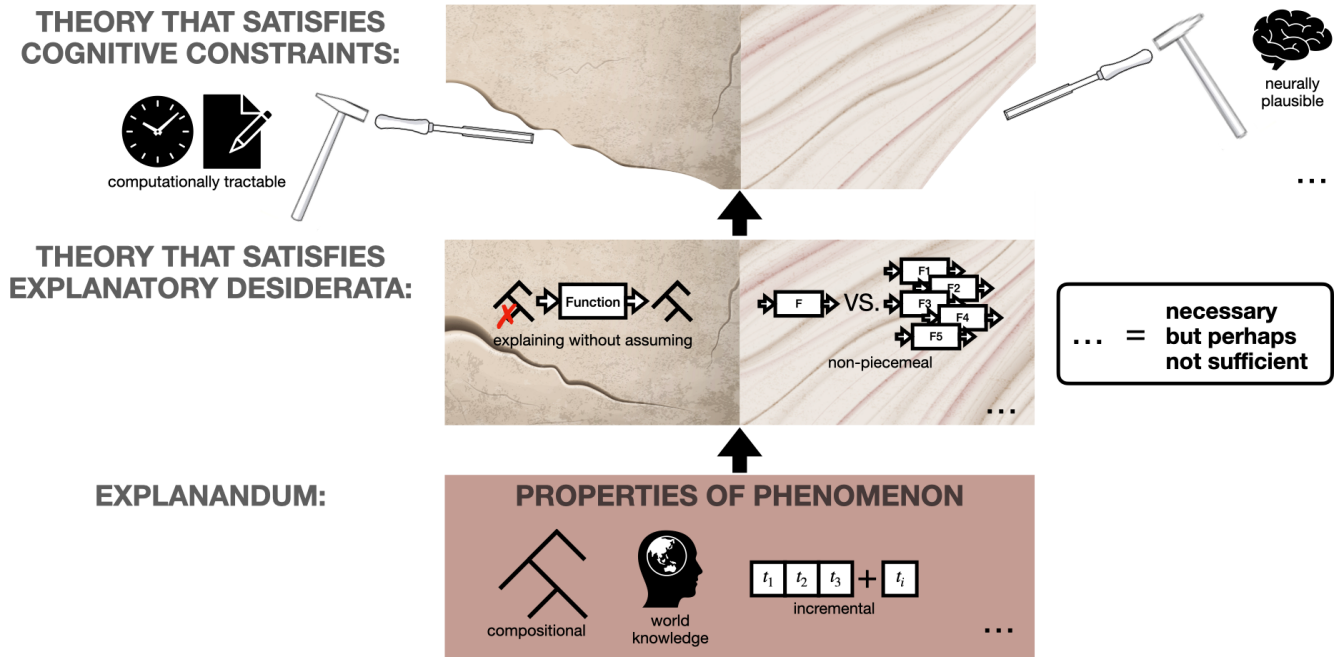
Figure 3: Illustration of how the explanandum and the constraints on the explanans relate to each other. The explanandum is characterised by three core properties. To explain this explanandum, a theory has to be able to capture these three properties (i.e., be *generatively sufficient*; Blokpoel, 2018). Evaluating the theory along the explanatory desiderata will help make sure that it really provides an explanation of the phenomenon of interest, and making sure the theory fits within the cognitive constraints will help make sure it can be plausibly realised by human cognition and the brain. Stone and tool images taken from freepik.com

a linear input sequence to a hierarchical, compositional representation. The latter is what requires explanation, and this computational process *itself* cannot be assumed to be part of the input domain, else we are assuming without explaining.

Explaining the role of world knowledge tractably

Blokpoel et al. (2019) present a computational-level model of how novel candidate hypotheses may be generated through *deep analogical inference*: where structured representations of knowledge are (iteratively) related to each other through analogy, in a way that allows for augmentation of structured representations (through projection of parts from one representation to its analogous representation). Blokpoel et al. (2019) highlight several necessary properties for this model, one of which is *isotropy*: That all knowledge is potentially relevant in the inference process (see also Blokpoel, 2015, chapter 1; and Fodor, 1983, part IV). The explanandum we focus on in this paper is related to the explanandum of Blokpoel et al. (2019) in the sense that flexibly coming up with plausible interpretations of neologisms probably requires coming up with *novel* structured representations based on the combination of linguistic knowledge and world knowledge that is activated by the incoming expression. In fact, this type of flexible interpretation of novel communicative signals is exactly the example that Blokpoel et al. (2019) use to illustrate their explanandum. The question they ask is: How are candidate hypotheses generated in abductive inference?

Where (a) plausible interpretation(s) of a novel communicative expression is an example of such candidate hypotheses. This raises issues for computational tractability, because if all world knowledge is potentially relevant, how can this component of the input domain be constrained? (See Section *Computational tractability* and Blokpoel et al., 2019, Section 4.2.)

Conclusion

Above, we worked out two examples of tensions that arise between different components of the force field that we identified in this paper. What we learn from these examples is that it is challenging even to satisfy one of the metatheoretical commitments that we put forward as important for explaining cognitive processes (i.e., the explanatory desiderata and cognitive constraints), while at the same time doing justice to each core property of the phenomenon (flexible linguistic inference) in its full capacity. Moreover, in the two examples above we limited ourselves to pairwise tensions between one property of the phenomenon and one metatheoretical commitment, but three-way or more-way tensions are also possible.

Other pairwise tensions that we did not have space to cover in this paper include: (i) explaining incremental comprehension in a non-piecemeal way (how to account for different levels of linguistic analysis?); (ii) explaining compositional comprehension in an neurally plausible way (compositionality requires symbolic processing—variable-value independence—while the brain excels at statistical and asso-

ciative learning; Martin & Baggio, 2020); and (iii) compositional comprehension and the involvement of world knowledge (how are world knowledge and linguistic knowledge integrated?). We encourage theoreticians to work out three- or more-way tensions between the different properties and constraints we put forward in this paper. To conclude, explaining flexible linguistic inference while satisfying these properties and constraints (Figure 3) poses a major challenge. The theory-development needed to solve this challenge, requires a keen awareness of the force field we exposed in this paper.

Acknowledgments

References

Adolfi, F., van de Braak, L., & Woensdregt, M. (2023). *From empirical problem-solving to theoretical problem-finding perspectives on the cognitive sciences.* OSF.

Adolfi, F., Wareham, T., & van Rooij, I. (2023). A Computational Complexity Perspective on Segmentation as a Cognitive Subcomputation. *Topics in Cognitive Science*, *15*(2), 255–273.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

Blokpoel, M. (2015). *Understanding understanding: A computational-level perspective*. Unpublished doctoral dissertation.

Blokpoel, M. (2018). Sculpting Computational-Level Models. *Topics in Cognitive Science*, *10*(3), 641–648.

Blokpoel, M., Wareham, T., Haselager, P., Toni, I., & van Rooij, I. (2019). Deep Analogical Inference as the Origin of Hypotheses. *The Journal of Problem Solving*, *11*(1), 1–24.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Guest, O. (2024). What Makes a Good Theory, and How Do We Make a Theory Good? *Computational Brain & Behavior*.

Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*, *6*(2), 213–227.

Hagoort, P. (2019). The meaning-making mechanism(s) behind the eyes and between the ears. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(20190301).

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438–441.

Hagoort, P., & van Berkum, J. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 801–811.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Kaushik, K. R., & Martin, A. E. (2022). *A mathematical neural process model of language comprehension, from syllable to sentence.* PsyArXiv.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, *38*(3), 245–294.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press.

Martin, A. E. (2016). Language Processing as Cue Integration: Grounding the Psychology of Language in Perception and Neurophysiology. *Frontiers in Psychology*, *7*(120).

Martin, A. E. (2020). A Compositional Neural Architecture for Language. *Journal of Cognitive Neuroscience*, *32*(8), 1407–1427.

Martin, A. E., & Baggio, G. (2020). Modelling meaning composition from formalism to mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(20190298).

McQueen, J. M. (2007). Eight questions about spoken word recognition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics*. Oxford University Press.

Partee, B. H. (1995). Lexical semantics and compositionality. In *Language: An invitation to cognitive science, Vol. 1, 2nd ed* (pp. 311–360). MIT Press.

Pylkkänen, L. (2020). Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(20190299).

Sedivy, J. C. (2007). Implicature During Real Time Conver-

sation: A View from Language Processing Research. *Philosophy Compass*, *2*(3), 475–496.

Slonimska, A., Özyürek, A., & Capirci, O. (2020). The role of iconicity and simultaneity for efficient communication: The case of Italian Sign Language (LIS). *Cognition*, *200*(104246).

Szabó, Z. G. (2004). Compositionality. In *Stanford Encyclopedia of Philosophy*.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, *268*(5217), 1632–1634.

van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, *32*(6), 939–984.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, *1*(3), 127–128.

van Rooij, I., & Blokpoel, M. (2020). Formalizing Verbal Theories. *Social Psychology*, *51*(5), 285–298.

van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (Eds.). (2019). *Cognition and intractability: a guide to classical and parameterized complexity analysis*. Cambridge University Press.

van Rooij, I., Guest, O., Adolfi, F. G., de Haan, R., Kolokolova, A., & Rich, P. (2023). *Reclaiming AI as a theoretical tool for cognitive science.* OSF.

van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional Communication: Computationally Easy or Difficult? *Frontiers in Human Neuroscience*, *5*.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.

Do Working Memory Constraints Influence Prediction in Verb-Final Languages?

Himanshu Yadav (himanshu@iitk.ac.in)
Department of Cognitive Science, Indian Institute of Technology Kanpur, India

Apurva (apurvajnu@gmail.com)
Business Communication, Indian Institute of Management Jammu, India

Samar Husain (samar@hss.iitd.ac.in)
Department of Humanities and Social Sciences, Indian Institute of Technology Delhi, India

Abstract

Prediction is argued to be a key factor in comprehending sentences in verb-final languages. The comprehender can predict the properties of the upcoming verb phrase using linguistic cues from the pre-verbal input. What are the constraints on prediction, beyond the ones posited by co-occurrence patterns in the language? We evaluate four models of verb prediction using data from a sentence completion study on Hindi. The model differs in their assumption of whether/how working memory constraints affect the prediction of the upcoming verb. The model comparison conclusively shows that working memory constraints do affect the prediction of the verb in Hindi. The results lead to a new insight into the underlying comprehension process: When the pre-verbal input is temporarily stored in memory, it probabilistically distorts to a non-veridical (or less accessible) memory representation, and this degraded representation of the context generates potentially faulty predictions of the upcoming verb.

Keywords: Sentence processing; prediction; working memory; representation distortion; encoding interference

Introduction

A key assumption in theories of sentence processing is that syntactic parsing happens incrementally by establishing links between related words in a sentence leading to a hierarchical structure (Frazier, 1987). The comprehender needs to integrate the currently parsed input with the upcoming sentence material. In verb-final languages, such as Hindi, prediction is argued to be a central feature of this incremental sentence processing (Konieczny, 2000; Vasishth & Lewis, 2006; Nakatani & Gibson, 2010; Husain, Vasishth, & Srinivasan, 2014; Levy & Keller, 2013). For instance, the cues from pre-verbal nominal modifiers can be used to predict certain features of the upcoming verb phrase. If the predicted linguistic features are compatible with the actual verb input, it is easier to integrate the verb with the previously built structure. Consequently, prediction can lead to a processing facilitation during incremental comprehension. What are the constraints on predictive processing in verb-final languages?

A well-formulated account of sentence processing — the lossy-context surprisal model — maintains that predictive expectations of the upcoming sentence material are subject to working memory constraints. The model assumes that currently encountered input, say words w_1 to w_{i-1}, which is temporarily stored in memory, undergoes information loss due to limited working memory such that certain features/words are deleted or inserted at constant rates. The comprehender uses this lossy input and their prior linguistic knowledge to generate predictions about the upcoming words, $w_i, w_{i+1}, \ldots$. As the information loss in the input increases, the predictions become more faulty. The model predicts that the less probable input with high working memory load is likely to produce more faulty predictions of the upcoming sentence material.

However, there is no empirical study that has directly tested this model prediction in a verb-final language. In this paper, we present a sentence completion study to test the effect of working memory constraints on verb prediction. The participants were asked to read sentences with three pre-verbal noun phrases where the nouns had either the same or distinct case markers. The null hypothesis is that case interference — that poses higher working memory load —- does not affect the rate of grammatical verb predictions.

We test four competing models that make different predictions about the distribution of verb prediction errors in these sentences. The first model is the null hypothesis model which assumes that the rate of verb prediction errors is solely determined by the conditional probability of an upcoming verb given the veridical representation of the pre-verbal input. The model draws its assumptions from the surprisal theory (Levy, 2008a; Hale, 2001). The second model — the memory interference model— assumes that case similarity causes representation degradation of the nouns making them less accessible in memory, and these less accessible nouns generate weaker/noisier predictions. The other two models — the lossy-context surprisal models — assume that when the nouns are stored in memory, the case markers are deleted/inserted at constant rates causing the comprehender to regress to most probable pre-verbal input. This leads to faulty predictions of the verb. We find that the lossy-context model whose information loss function was estimated empirically shows the best predictive performance among the models considered here. However, this model was not distinguishable in its performance from the memory interference model that assumed representation degradation of nouns causes imprecise predictions. Overall, the results suggest that prediction during sentence comprehension is constrained by probabilistic distortion of representations stored in memory.

We first present the four models of verb prediction. Next, we quantify the predictive performance of the models using data from a sentence completion experiment in Hindi. We then discuss the broader implications and conclude.

Computational models of verb prediction

We implement four models of verb prediction that differ in their assumptions about how working memory constraints affect the prediction of the upcoming linguistic items. The first model is the null hypothesis model and assumes no explicit effect of working memory constraints on prediction. The second model assumes that working memory constraints cause encoding interference when similar case-marked nouns are stored in memory, which in, turn affects the quality of prediction. The remaining two models assume that actual pre-verbal input distorts to a lossy representation such that certain case markers are deleted or exchanged probabilistically and these lossy representations and the prior linguistic knowledge is used by the comprehender to make potentially faulty predictions. We specify the assumptions and prior predictions of these four models for the following set of pre-verbal input sentences in Hindi.

(1) a. ruchi-ko   sumit-ne   priya-ko   …
         Ruchi-*ACC* Sumit-*ERG* Priya-*ACC* …

     b. ruchi-se   sumit-ne   priya-se   …
         Ruchi-*ABL* Sumit-*ERG* Priya-*ABL* …

     c. ruchi-ko   sumit-ne   priya-se   …
         Ruchi-*ACC* Sumit-*ERG* Priya-*ABL* …

In the above example, conditions (1a) and (1b) are the case interference conditions, where two of the nouns posses similar case markers while condition (1c) is no interference condition. We collect sentence completion data for the conditions (1a –c) and compare the performance of the four models on the observed rate of grammatical verb completions in each condition.[1]

The surprisal model

The surprisal model (Levy, 2008a; Hale, 2001) assumes that given a sentential context say words w_1, w_2, w_{n-1}, the probability of encountering an upcoming word, say w_{nj}, is assigned by a distribution of conditional probabilities over all possible continuations given the input $w_1, w_2, ..., w_{n-1}$.

$$w_{nj} \sim p(w_n | w_1, w_2, ..., w_{n-1}) \qquad (1)$$

For the contexts $(1a - c)$, the conditional probability of seeing a non-finite verb or a causative verb can be derived from the corpus. The rate of grammatical verb completions in condition (1a) and (1b) will be given by the conditional probability of seeing a non-finite verb completion, and in condition (1c), it will be given by the sum of conditional probabilities of seeing the non-finite verb and the causative verb. As the conditional probabilities for the surprisal model are computed from a corpus of grammatical sentences, the rate of grammatical verb continuations in all these sentences are close to 1. Thus, effectively the surprial model predicts no difference between the three conditions (1a), (1b) and (1c).

---

[1]Here, the label *ACC* means Accusative case marker, the label *ERG* means the Ergative case marker, and *ABL* means the Ablative case maker.

The memory interference model

The model assumes that encoding interference —the interference due to nouns with similar features stored in memory— adversely affects the prediction of the upcoming linguistic input. We implement this assumption using a simple memory interference model based on the feature overwriting theory of (Oberauer & Kliegl, 2006). The model assumes that

- A noun stored in memory maintains a degree of representation/accessibility determined by the number of feature units it shares with other pre-verbal nouns. As the number of shared feature units increases, the representation degradation increases.

- The representation quality of the pre-verbal noun determines the quality of the prediction: The rate of grammatical verb completions is a function of the activation of the pre-verbal nouns in memory.

The model was implemented as follows. The activation of a pre-verbal noun (in the feature layer of the memory) is a function of interference arising from similar features stored in memory. The activation of the noun i that shares K feature units with the other nouns is given by:

$$A_i = (1) - r \frac{K}{2n}^{n-1} \qquad (2)$$

where n is the total number of feature units for the noun i, r is the rate of feature decay due to overwriting — the larger the value of r, the higher the representation degradation of the noun.

The probability of correct (grammatical) verb completions is a logistic function of the activation of the subject noun A_i,

$$P_{gram} = \frac{1}{1 + e^{-\frac{(A_i - a_0)}{T}}} \qquad (3)$$

where T is the noise factor that determines how strongly the noun's activation level impacts the quality of verb prediction. As the noise factor T grows larger and larger, the rate of grammatical continuations goes towards chance-level performance.

Figure 1 shows the prior predictions of the model for the three conditions shown in Example 1. The model predicts a relatively small rate of grammatical verb continuations in the case interference conditions 1a and 1b compared to condition 1c.

The lossy-context surprisal models

The lossy-context surprisal model (Futrell, Gibson, & Levy, 2020) assumes that the comprehender has access to only a non-veridical, lossy representation of actual pre-verbal input and uses this lossy representation and their prior linguistic knowledge to make probabilistically faulty predictions about the upcoming verb. For example, consider the sentence 1a *Ruchi-ko Sumit-ne Priya-ko ...*, the input here is

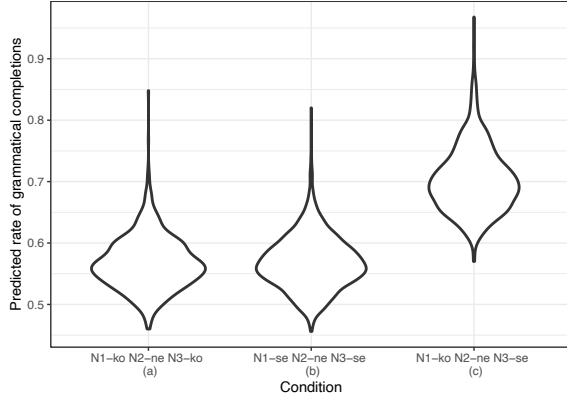$I = N1-ko \quad N2-ne \quad N3-ko$ …

Figure 1: The mean rate of grammatical completions predicted by the memory interference model in the three conditions. In conditions (a) and (b), the model predicts lesser grammatical completions due to interference arising from ko and se markers respectively.

where N represents a noun, $-ko$ represents an accusative marker, and $-ne$ represents an ergative marker in Hindi.

The input I gets distorted to a possible memory representation r_i such that the first noun gets deleted with probability d or the case marker on the first noun gets exchanged with an ablative case marker with probability e.[2]

The following memory representations are possible,
$$r_1 = N1-ko \quad N2-ne \quad N3-ko$$
$$r_2 = N2-ne \quad N3-ko$$
$$r_3 = N1-se \quad N2-ne \quad N3-ko$$

The model further assumes that the comprehender reconstructs a set of possible, true pre-verbal contexts from their memory representation r_i based on their prior linguistic knowledge and their uncertainty about the degree of information loss in the system. We can derive the conditional probability of producing a grammatical verb completion $P(V_{gram}|r_i)$ by marginalizing out all possible true contexts $c_1, c_2, ..., c_n$

$$P(V_{gram}|r_i) = \sum_{j=1}^{n} P(V_{gram}|c_j)P(c_j|r_i) \quad (4)$$

where $P(V_{gram}|c_j)$ is the conditional probability of seeing the verb given a possible true context c_j. We can derive the probability $P(c_j|r_i)$ up to proportionality using Bayes' rule,

$$P(c_j|r_i) \propto \mathcal{L}(c_j|r_i)P(c_j) \quad (5)$$

where $P(c_j)$ represents the prior probability of seeing the representation c_j in the corpus, and $\mathcal{L}(c_j|r_i)$ is the likelihood of obtaining the memory representation r_i from a possible true representation c_j.

---

[2]The other representations are possible due to case deletions and insertions. For example, the case marker on the first noun can get deleted probabilistically. However, such representations are very unlikely (ungrammatical) to occur in Hindi. Therefore, we ignore them for simplicity and well-constrained model specifications for now.

Based on equations 4 and 5, we can rewrite the conditional probability of seeing a grammatical verb completion as

$$P(V_{gram}|r_i) \propto \sum_{j=1}^{n} P(V_{gram}|c_j)\mathcal{L}(r_i|c_j)P(c_j) \quad (6)$$

The likelihood function $\mathcal{L}(c_j|r_i)$ is called the lossy memory encoding function: the likelihood that a true representation c_j gets distorted to memory representation r_i given a deletion rate d and case exchange rate e.[3]

$$r_i|c_j \sim Memory(d, e) \quad (7)$$

where $Memory(d, a)$ is the lossy memory encoding function, d is the rate of deleting the first noun, and e is the rate of exchanging the case marker on the first noun.

We implement two versions of the lossy-context surprisal model that differ in their assumptions about the lossy memory encoding function:

1. **Deletion error model:** The model assumes that the input can distort to a memory representation such that the first noun gets deleted with a probability d (deletion error). The exchange error rate e is assumed to be zero in this model. For example, when the comprehender stores the input $N1-ko \, N1-ne \, N1-ko...$ in memory, the noun $N1i$-ko can get deleted in some trials to produce $N1-ne \, N1-ko...$.

2. **Deletion-and-exchange-error model:** The model assumes that the input can distort to a memory representation such that the first noun gets deleted with a probability d (deletion error) and the case marker on the first noun can get changed to another case marker with probability e (case exchange error). For example, in the input $N1-ko \, N1-ne \, N1-ko...$, the ko marker on the first noun can change to se marker in some trials to produce $N1-se \, N1-ne \, N1-ko...$.

Do we have any experimental evidence that both the noun deletion errors and the case exchange errors occur in these sentences? The above implementation of the model allows a relatively high modeler's degree of freedom. A principled way will be to estimate the nature of the information loss function experimentally. For example, we can experimentally estimate whether both deletion and case exchange occur or only one of them does.

We did a rating study to estimate to rates of deletion and exchange error rates. The participants were asked to rate sentences like A–C shown in Table 1. Among these sentences, A and B can be treated as grammatical after a noun deletion error or a case exchange error, and sentence C cannot become grammatical due to such information loss mechanisms.

---

[3]The same memory function also underlies $P(r_i|I)$: the probability of generating a memory representation r_i from the observed linguistic input I. The function $P(r_i|I)$ represents the experimenter's uncertainty about the memory representation formed by the comprehender and the function $\mathcal{L}(r_i|c_j)$ represents the comprehender's uncertainty about the true intended input.

The sentences were presented word-by-word and participants were asked to judge the acceptability of the sentence (on a 1-100 scale) after the last word disappeared. If the participants make noun deletion and case exchange errors, sentences A and B should get higher acceptability ratings than sentence C. Moreover, if deletion error rates are higher than the exchange error rates, sentence *A* should get the highest acceptability ratings.

The acceptability rating results are shown in Table 2. We observe that the deletion and exchange errors do occur as the ratings of A and B are significantly higher than the C. And, the rate of noun deletion and case exchange errors is approximately the same (27-29% acceptability in B and C). We use these ratings to estimate the rate of deletion and exchange errors. For example, the deletion error rate can be estimated as the rating of sentence A minus the rating of baseline ungrammatical sentence C divided by the rating scale (100) and so on. This is an approximate and indirect measure of error rates under the assumption that sentences A and B are perceived more grammatical compared to C due to a corresponding rate of deletion and exchange errors in these sentences. The empirically estimated error rates suggest that the second lossy model that assumes both deletion and exchange errors is a better approximation of the underlying memory encoding function.

We specify the following priors on the deletion rate *d* and the exchange error rate *e* informed by their empirical estimates:

$$d \sim Normal_{lb=0.1, ub=0.5}(0, 0.2) \quad (8)$$

$$e \sim Normal_{lb=0.1, ub=0.5}(0, 0.2) \quad (9)$$

where $lb = 0.1$ is the lower bound on the deletion rate and exchange rate values. The parameters *a* and *e* represent the rate of information loss when the linguistic input is stored in memory.

Table 1: Sample items for the '*ko ne ko*' prefix conditions in the acceptability experiments. Note: the items in condition A can be treated as grammatical if the light gray element (*N1-ko*) is ignored. Similarly, the items in condition B can be treated as grammatical only if the light gray colored element (*ko*) is treated as a *se*; otherwise these items are ungrammatical. Condition C is the baseline condition and cannot be made grammatical through the mechanisms stated above. ne=Ergative case-marker, ko=Accusative case-marker, se=Ablative case-marker.

Condition	Sample item
A (N1 deletion)	N1-ko N2-ne N3-ko kitaab padhne ko kaha
	'asked to read the book'
B (N1 case-exchange)	N1-ko-se N2-ne N3-ko kitaab lene ko kaha
	'asked to take the book'
C (Inherently ungrammatical)	N1-ko N2-ne N3-se so gaya
	'slept'

Table 2: The acceptability rating (on a scale of $1 - 100$) of ungrammatical sentences of type A, B, and C, where A are the sentences that become grammatical if the first noun is deleted, B are the sentences that become grammatical if the case marker on the first noun changes to another marker, and C sentences are inherently ungrammatical.

	Condition	Acceptability rating
A	First noun deletion can make it grammatical	**29.8** (sd=5.8)
B	Case exchange on the first noun can make it grammatical	**26.6** (sd=5.2)
C	Baseline ungrammatical	**6.1** (sd=3.4)

(1) The deletion error model
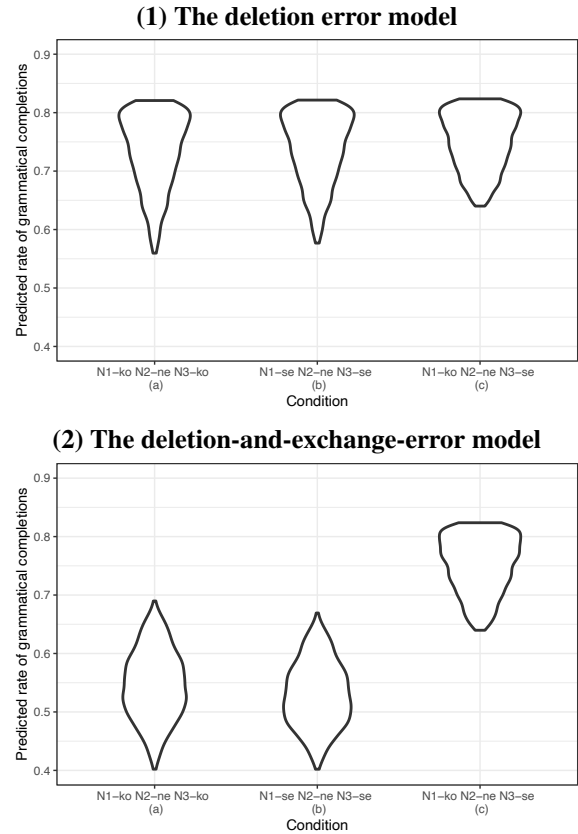
(2) The deletion-and-exchange-error model

Figure 2: The mean rate of grammatical completions predicted by the lossy-context models, the deletion-only error model and the deletion as well as case exchange error model. In conditions (a) and (b), the deletion-and-exchange error model predicts lesser grammatical completions due to interference arising from ko and se markers respectively.

Figure 2 shows the prior predictions of the two lossy-context models. The deletion error model predicts that the rate of grammatical completion is slightly low in all three conditions but there is no considerable difference across the conditions. The deletion-and-exchange error model predicts the rate of grammatical completion is considerably lower in conditions (a) and (b) where the nouns share a case marking feature compared to condition (c).

Model comparison

We compare the predictive performance of the above four models of verb prediction using data from a sentence completion study in Hindi. First, I describe the sentence completion task and the data and then show the quantitative evaluation of the completing models.

Data

We did a sentence completion study (Taylor, 1953), where incomplete items described in (2) were displayed on a computer screen using the centered self-paced reading paradigm. The participants were instructed to complete the sentences to make them meaningful. The sentences comprised of three nouns with three case markers depending on the condition, which could be *ne* (Ergative case), *se* (Ablative/Instrumental case), or *ko* (Accusative case). The first and the third nouns possessed identical features (+animate, +female) in conditions (a) and (b). We used three combinations of case markers: (a) N1-ko N2-ne N3-ko, (b) N1-se N2-ne N3-se, and (c) N1-ko N2-ne N3-se. 12 items were prepared for each case-marker combination; 36 native speakers of Hindi participated in this experiment.

(2)　a.　**N1-ko N2-ne N3-ko**

　　　ruchi-ko　　sumit-ne　　priya-ko　　…
　　　Ruchi-*ACC* Sumit-*ERG* Priya-*ACC* …

　　b.　**N1-se N2-ne N3-se**

　　　ruchi-se　　sumit-ne　　priya-se　　…
　　　Ruchi-*ABL* Sumit-*ERG* Priya-*ABL* …

　　c.　**N1-ko N2-ne N3-se**

　　　ruchi-ko　　sumit-ne　　priya-se　　…
　　　Ruchi-*ACC* Sumit-*ERG* Priya-*ABL* …

We response coded the sentence completions such that all grammatical completions were coded as '1' and ungrammatical completions as '0'. Table 3 shows the rates of grammatical completions across the three conditions. We find that the grammatical completions for the conditions with similar case markers, i.e., for conditions (a) and (b), are around 50%. The condition with distinct case-markers is significantly higher at 75%. A mixed-effect logistic regression analysis shows that the rate of grammatical completions in condition (c) is significantly lower than in condition (a) (p<0.0001) and condition (b) (p<0.0001).

Table 3: The rate of grammatical completions for the three combinations of case-markers.

Condition	Rate of grammatical completions
(a) *N1-ko N2-ne N3-ko*	0.52
(b) *N1-se N2-ne N3-se*	0.54
(c) *N1-ko N2-ne N3-se*	0.75

Quantitative model comparison results

The models were evaluated using leave-one-out cross-validation on the data from the sentence completion task. We use the difference of expected log pointwise predictive densities, $\Delta \widehat{elpd}$, measure to assess the relative predictive accuracy of the models. The higher the $\Delta \widehat{elpd}$ for two models, the larger the difference between their predictive performances. We find the following key results:

1. All models that assume representation degradation/distortion of the pre-verbal input perform considerably better than the surprisal model, which assumes no effect of memory constraints on prediction (all $\Delta \widehat{elpd} > 300$).

2. The best predictive performance is shown by the lossy-context surprisal model that assumes both deletion and case exchange errors ($\widehat{elpd} = -836, SE = 10.8$).

3. The memory interference model is comparable to the lossy-context surprisal model that assumes both deletion and case exchange errors ($\Delta \widehat{elpd} = 1.5(SE = 2.2)$).

4. The memory interference model shows superior performance compared to the lossy-context surprisal model that assumes only the noun deletion errors. ($\Delta \widehat{elpd} = 5.7(SE = 2.1)$).

The results suggest that the verb completion data can be explained by either – (i) **a memory interference model of verb prediction** which assumes that the representation of the nouns with similar case marking degrades in memory due to a feature overwriting process; these less accessible nouns cause difficulty in predicting the correct verb continuations; or (ii) **a lossy-context surprisal model** which assumes that pre-verbal nouns stored in memory undergo distortions due to deletion and case exchange errors, consequently, producing faulty predictions of the verb. Taken together, the results indicate that probabilistic distortion of the pre-verbal input stored in memory constrains the prediction of the upcoming verb.

The reason behind the inferior performance of the deletion-error-only model is that the first noun deletion errors produce locally coherent structures that have large prior probabilities in all three conditions. These non-veridical memory representations containing two marked nouns will predict transitive verbs which are equally ungrammatical continuations for all three conditions. Thus, the model with the deletion error noise ignores the possibility of any specific error due to case interference in conditions (a) and (b).

In contrast, the presence of case exchange errors predicts a special cost in interference conditions. For example, in condition (a), the *ko* marker on the first noun can change to *se* marker with a probability *e*. In these cases, this non-veridical memory representation has a higher prior probability and will predict either a causative verb or a non-finite verb. Since causative verbs are not compatible with the actual context in conditions (a) and (b), we get a large number of un-grammatical continuations, compared to condition (c) where causative verbs are compatible with the context. Importantly, noun deletion errors are necessary to explain the low rate of grammatical continuations in all three conditions. Hence, a lossy-context surprisal model that assumes both noun deletion errors and case exchange errors best explains the data.

Overall, these modeling results support the hypothesis that memory-based constraints, more specifically representation distortion of the linguistic input, affect the prediction of the upcoming linguistic material in verb-final languages.

Discussion

Is the prediction of the upcoming sentence material affected by working memory constraints on the maintenance of the previous context? To answer this question, we implemented four models of verb prediction, the suprisal model, the memory interference model, and two lossy-context suprisal models. The surprisal model implements the null hypothesis that the prediction of the verb is only constrained by its statistical co-occurrence with the pre-verbal context and there is no explicit influence of working memory on prediction. The memory interference model assumes that pre-verbal nouns with similar features are difficult to encode and maintain in memory which in turn affects the quality of prediction of the upcoming verb. Finally, the lossy-context surprisal models assume that the pre-verbal input distorts to a lossy representation due to working memory constraints such that certain nouns and case markers are deleted or exchanged at constant rates. The comprehender uses a lossy memory representation of the actual context and their prior linguistic knowledge to make predictions about the upcoming verb. The models were evaluated on sentence completion data from Hindi, a verb-final language.

The model comparison revealed two key insights: (i) The assumption that the pre-verbal input undergoes probabilistic representation degradation is necessary to explain the verb prediction data, as all the models under this assumption performed better than the null hypothesis model, which assumed no distortion of the pre-verbal input; and (ii) The distortion of the pre-verbal input occurs either due to deletion and insertion of features or due to a feature overwriting process when nouns share certain features during encoding.

The results indicate that prediction is affected by working memory constraints: When a sentential context is temporarily stored in memory, it undergoes probabilistic representation distortion due to working memory limitations, and consequently, it generates faulty predictions of the upcom-ing sentence material. The results are important for theories of sentence processing because they conclusively show that prediction —which is viewed as an important factor in the comprehension process— is constrained by working memory limitation. As the working memory load increases on temporarily stored linguistic input, the prediction of upcoming linguistic items becomes noisier and faulty. For comprehension in verb-final languages, where prediction is argued to be a central processing strategy, the results imply that prediction of the verb is robust and useful only when the pre-verbal is simple. A complex pre-verbal input is more likely to distort to a non-veridical representation in memory causing faulty prediction of the upcoming verb.

Another major implication is for studies that invoke prediction as an explanation in contrast to the working memory explanation. For example, the anti-locality effects —where an increase in the number of pre-verbal modifiers causes processing facilitation at the verb— observed in verb-final languages are explained by prediction-based accounts; the prediction of verb gets better with increased pre-verbal material leading to facilitation at the verb (Konieczny, 2000; Vasishth & Lewis, 2006; Husain et al., 2014). The prediction explanation is often invoked in contrast to the working memory explanation, e.g., strong predictive expectations in verb-final languages override certain working memory constraints (Husain et al., 2014). However, our results show that prediction is not independent of working memory constraints; in fact, it is drastically influenced by working memory.

Finally, this study lays a framework for systematic evaluation of the lossy-context theories of sentence processing (Futrell et al., 2020; Levy, 2008b). The lossy-context models typically allow a lot of modeler's degree of freedom in specifying the lossy memory encoding function. For example, the framework allows to freely assume which kind of errors cause the information loss and which types of words or morphemes are lost or changed with time. As a consequence, the model is allowed to generate unconstrained predictions for a given comprehension task. To constrain model predictions, the memory encoding function must be restricted in its assumptions. A systematic way to achieve this is to estimate what kind of memory encoding errors do occur during comprehension. Our work is the first attempt to empirically estimate the nature of the lossy encoding function and to use these estimates to constrain the model predictions.

The current work reveals new insights about the top-down predictive processes during sentence comprehension: The prediction of the upcoming sentence material is modulated by probabilistic distortion of the previous linguistic context stored in memory. To our knowledge, this is the first empirical study that shows that prediction is constrained by working memory limitations. Our work contributes to understanding how working memory constraints and predictive processes interact and how these two factors can be integrated to build a unified theory of sentence processing.

References

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *The psychology of reading* (Vol. 12, pp. 559–586). Hillsdale, NJ: Erlbaum.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814. doi: https://doi.org/10.1111/cogs.12814

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from hindi. *PloS One*, *9*(7), e100986.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of psycholinguistic research*, *29*(6), 627–645.

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 234–243). Association for Computational Linguistics.

Levy, R., & Keller, F. (2013). Expectation and locality effects in german verb-final structures. *Journal of memory and language*, *68*(2), 199–222.

Nakatani, K., & Gibson, E. (2010). An on-line study of japanese nesting complexity. *Cognitive Science*, *34*(1), 94-112.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of memory and language*, *55*(4), 601–626.

Taylor, W. L. (1953). cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 767–794.