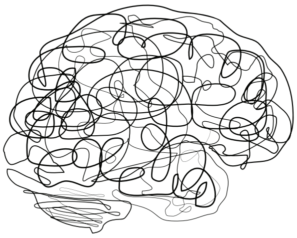


PROCEEDINGS OF THE 18TH SWECOG CONFERENCE

Göteborg 2023

5 - 6 October

Editors

Pierre Gander
Linus Holm
Erik Billing

SweCog
SWEDISH COGNITIVE
SCIENCE SOCIETY

Editors:
Pierre Gander, University of Gothenburg, pierre.gander@gu.se

Linus Holm, Umeå University, linus.holm@umu.se

Erik Billing, University of Skövde, erik.billing@his.se
ljud www.swecog.se

Skövde University Studies in Informatics 2023:1
ISBN 978-91-989038-0-5
ISSN 1653-2325

PUBLISHED BY THE UNIVERSITY OF SKÖVDE

Preface

Welcome to the annual conference of the Swedish Society for Cognitive Science in Göteborg, SweCog 2023. The impact of AI widely discussed in society today is reflected by a substantial fraction of the conference contributions this year, targeting challenges in interactions with artificial intelligence. The challenges raised include how to develop trust in AI and how to explain its operation to a human user. Addressing these applied research questions require revisiting more fundamental questions in the cognitive sciences such as what trust and explanation is, what information requirements they impose, and constraints in human decision-making and learning. Happily, the contributions this year cover this breadth of enquiry. Perhaps the rise of AI offers a mirror that help us develop a better view of human reasoning, widely construed? Yet all may not be well in this technological development – what are the risks imposed by relying on artificial intelligence as curator or source of information? This topic is directly targeted by the panel discussion *“what is left to the human mind when machines do the thinking?”*. Please join us at the conference to participate in the development of the emerging answers to these both old and new questions.

Conference Programme

Thursday October 5th

12:30 — 13:00	*Registration*
13:00 — 13:50	Invited speaker — **Karin Jensen** *Predictions of relief: the science of the placebo effect*
13:50 — 15:15	Oral presentation session 1 with Amandus Krantz [p. 11], Katie Winkle [p. 13], and Ilaria Torre [p. 79]
15:15 — 15:45	*Coffee break*
15:45 — 16:50	Oral presentation session 2 with Melina Tsapos [p. 21], Alexander Berman [p. 5], and Andreas Chatzopoulos [p. 6]
16:50 — 18:00	**Poster session with elevator pitch**, hosting Ismael Albutihe [p. 5], Anton Smedberg et al. [p. 6], Philip Gustafsson [p. 7], Erik Hallberg & Ludwig Lundstedt [p. 7], Azadeh Karamali [p. 8], Erik Lagerstedt [p. 9], Kajsa Nalin & Erik Lagerstedt [p. 10], Anders Persson [p. 10], Samantha Stedtler [p. 11], Nanna Strid et al. [p. 12], Franziska Babel et al. [p. 17], Victor Nyberg [p. 31], William Hedley Thompson [p. 37], and Shuren Yu [p. 83]

Friday October 6th

09:00 — 09:50	Invited speaker — **Pär-Anders Granhag** *The detection of lies: Emotional vs. cognitive approaches*
10:10 — 11:50	Oral presentation session 3 with Mattias Forsgren [p. 27], Linus Holm [p. 8], Maybí Morell Ruiz [p. 9], Mohammad Hossein Heydari Beni [p. 47], and Mattias Rost [p. 73]
11:50 — 13:00	*Lunch*
13:00 — 14:00	Oral presentation session 4 with Betul Tolgay & Oskar MacGregor [p. 13], Simon Skau [p. 61], and Dániel Pénzes [p. 53]
14:00 — 15:00	Invited speaker — **Virginia Dignum** *What is Responsible AI and why should you care*
15:00 — 15:40	Panel discussion on AI with Prof. Virginia Dignum, Prof. Jonas Ivarsson, and Prof. Olle Häggström, moderated by Assoc. Prof. Linus Holm: *What is left to the human mind when machines do the thinking?*
15:40 — 15:50	Conference closing

Abstracts

Sense of Agency and Automation

Ismael Albutihe

Shool of Bioscience, University of Skövde

Technological evolution has resulted in complex automated systems present in various tools and devices, such as self-driving cars and autopilot systems. This progress has led to a new type of interaction known as ”Human-Robot Joint Action” or ”Human-AI Interaction.” The effects of this interaction on the users' sense of agency (SoA; i.e., the sense of generating an action outcome) with these automated tools are not well understood. An altered sense of agency might not only impact the users' experiences but also their ability to successfully control task outcomes. This systematic review examines how automated tools at different levels of automation affect SoA. We conducted a systematic review using Scopus and MEDLINE EBSCO databases, which yielded eight relevant articles for review. Preliminary findings suggest that as tools become more automated, there is a corresponding decrease in the SoA experienced by human users. Interestingly, the impact of different automation levels on human SoA appears to depend on the nature of the task at hand. However, this area of studying the psychological impact of automated tools is still in its early stages. Further research is needed to fully understand and draw definitive conclusions about the nuances in Human-AI Interaction and its effects on SoA.

Counterfactual reasoning capabilities of GPT: Preliminary findings

Alexander Berman and Christine Howes

Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg

Recently, there has been a large interest in large language models (LLMs) such as GPT and their ability to engage in human-like dialogue and use commonsense reasoning. We experimentally investigate specific aspects of these abilities, namely counterfactual reasoning and explanations. These abilities are particularly important when using LLMs to assist high-stake decisions and assessments such as credit approval or medical diagnostics. For example, if a loan applicant is denied credit, a counterfactual explanation conveys the conditions under which the credit would have been granted. By injecting a decision-making algorithm into the model's prompt and systematically probing and annotating responses for carefully chosen inputs, we study potential patterns in GPT's selection of counterfactual examples. Preliminary results indicate that when GPT 3.5 provides counterfactual explanations, it does not consider causal relations between variables in a way that one would expect from a model with strong commonsense reasoning capabilities. We discuss potential implications of these results for real-world applications and future research.

Cognitive explanations and Large Language Models

Andreas Chatzopoulos

Dept. of Applied Information Technology, University of Gothenburg

There has been an ongoing debate in the Cognitive Science community regarding what type of cognitive explanations should be employed, and to what degree the can be considered scientific explanations. The debate has centered around the supposedly growing prevalence of dynamical explanations at the expense of more traditional, mechanistic explanations. However, regardless of the underlying type of cognitive explanation, it is always possible to use them as a basis for models and simulations, important tools for Cognitive Science.

But where do Large Language Models fit into this picture? They are undeniably models, but of what kind? They are not based on cognitive explanations and are not meant to be realistic representations of the brain. Nevertheless, they can be an important tool for understanding cognitive phenomena.

To make theoretical sense of all this, I will suggest a philosophical analysis of Large Language Models to show how they are related to more traditional models and simulations, meant to represent reality. To this end, I will use a philosophical distinction between different types of scientific explanations, and use it to show that Large Language Models can be seen as models based on what some philosopher calls ”how-roughly-explanations”.

Art by Man vs. Machine: The effect of human versus AI as attributed creator when rating artworks

Anton Smedberg, Filip Kockendal, and Pierre Gander

Dept. of Applied Information Technology, University of Gothenburg

Art has been a means of human expression for centuries, but the rise of artificial intelligence (AI) has brought new opportunities for artistic creation. However, concerns about the quality and authenticity of AI-generated art compared to human-created art have emerged as AI-generated art gains popularity. This study aims to investigate whether a negative bias against AI-generated art exists. In an experiment, which was designed as a between-groups study with the conditions AI-priming and human-priming, 233 participants were asked to evaluate 10 artworks based on four criteria (liking, beauty, uniqueness, and meaning). Both groups were presented with the same set of artworks, but given information that they were created by humans or AI, respectively. Participants who were informed that artworks were created by AI gave lower ratings than those who were informed a human artist created them. The main factors driving this difference were the criteria of ”unique” and ”beautiful,” which received lower ratings from the AI-primed group. The results also indicate that surrealist artworks received lower ratings when AI was attributed as the creator. Overall, the results implies an existing negative bias towards AI-generated art, despite the fact that AI is becoming an increasingly integrated part of people’s lives.

The voice of accuracy

Philip Gustafsson

Department of Psycology, Stockholm University

Eyewitnesses remembering incorrect facts poses a problem within criminal law and can lead to wrongful convictions. Finding indicators of correct memory recall has the potential to reduce injustice. In the current work, we show that the acoustical cues with which eyewitnesses speak can help us understand if they recall correctly or not. Relatedly, we examine to what extent people can detect these indicators of accuracy. Here, we show that observers can judge whether a testimony is correct or not with above-chance-accuracy when listening to eyewitnesses’ voices. This finding replicates across observers who are native speakers of the testimony language and observers from other nations who do not understand the testimony language. This suggests that there may be universal cues to accuracy, and opens a new avenue of research that may strengthen social justice by minimizing future wrongful convictions.

Agency of others: The intentional binding paradigm in observed actions

Erik Hallberg and Ludwig Lundstedt

Shool of Bioscience, University of Skövde

Sense of agency (SoA) is defined as the subjective experience of being in control of one’s own actions. This attribution of control underpins all voluntary action and is thought to be a critical aspect of the self. SoA is conceptualized as being self-specific, yet a number of studies have reported agency during the observation of other-generated actions. Our systematic review sought to address whether intentional binding (IB), a proxy of SoA, can be found during observation of other-generated actions. IB refers to the observation that in voluntary actions, action and outcome are perceived as being closer in time. In our systematic review, we have identified six articles that examined the experience of agency during action observation. These studies found that IB was present in different non self-related contexts which highlights the flexible nature of SoA. Most importantly, we concluded that IB can and does occur during the observation of other-generated actions. We theorize that social influence might have an effect on IB in both human- and robot-observations. The fact that IB can be found during observed other-generated actions raises questions for contemporary theories of SoA.

Information source credibility predicts curiosity in trivia fact learning

Linus Holm, Josef Vestin, Hanna Ebbvik Ivars and Felix Thiel

Department of Psychology, Umeå University

Curiosity has been suggested to reflect a drive for learning. As a rational signal for learning opportunity, curiosity should contain an assessment of the reliability of the information source. It then follows that more trusted sources should also instill more curiosity and learning. We tested these hypotheses in an experiment involving 23 student participants where we randomly assigned zoology trivia questions and correct answers to one of three different claimed sources (Encyclopedia Britannica, Wikipedia and Reddit). Participants were told that we estimated the correctness rate of the sources to .99, .90 and .75, respectively. Participants rated their curiosity for the answer to 100 questions, read the answers, and then took a retest on the items after a few minutes delay. We found that indicated answer source reliability significantly affected curiosity ratings yielding an average of .56z,.22z and -.78z, respectively. Memory performance mirrored the results numerically but not statistically reliably due to ceiling effects. Taken together, our results suggest that the perceived source reliability directly affects curiosity. If a simple manipulation of source trust such as this substantially impact curiosity and potentially learning, then deceptive targeting of source credibility might substantially misguide human willingness to learn.

The Evolutionary Origins of Consciousness

Azadeh Karamali

Shool of Bioscience, University of Skövde

Unanswered questions about the evolutionary origins and distribution of consciousness among living organisms persist. This review study aims to shed light on these age-old questions by examining the literature on evolutionary approaches to the fundamental concept of phenomenal consciousness. In alignment with the 'Cambridge Declaration on Consciousness,' this study introduces three recently developed theories, with a particular emphasis on examining one model. The 'Cellular Basis of Consciousness' (CBC) is a reductionist, cellular-based model that posits sentience in all organisms, from unicellular life forms to humans. Another theory, rooted in neuroevolutionary arguments, is 'Neurobiological Naturalism.' It suggests that consciousness first emerged during the Cambrian period, approximately 550 million years ago, and includes vertebrates, arthropods, and cephalopods as conscious animals. The primary focus of this study is the 'Unlimited Associative Learning' (UAL) framework, which employs a novel method to establish a transition marker as an indicator of consciousness. While the UAL framework shows promise for tracing the evolution of consciousness, it also has notable limitations. Nevertheless, the literature review indicates that UAL, as an innovative framework, holds the potential to initiate fruitful research programs. Rather than providing definitive answers, it can be regarded as a significant starting point for unraveling the origin of consciousness.

Pragmatic Existentialism as Framing for Cognition

Erik Lagerstedt

Interaction Lab, School of Informatics, University of Skövde

When experts describe what cognition is, there are typically some reoccurring concepts, for instance learning, remembering, perceiving, reasoning, imagining, and making decisions. However, what terms are emphasised (and what the nature of the corresponding concepts are) largely depend on what cognitive science paradigm the respective experts rely on and what related fields of study they come from. There is still no strong consensus regarding any single definition of cognition.

That said, it is hard to argue against existence as a prerequisite of cognition. On the one hand, such a broad claim might not be particularly useful since little is excluded by it. On the other hand, taking that claim seriously can help frame cognition by attaching it to the study of the nature of existence.

Taking a pragmatic stance (as in emphasising practice and utility of theories) to existentialism (where themes such as the dread of facing an unintelligible and uncaring universe are typically central) could ground cognition as something meaningful, since cognition could be considered an important tool for managing one's existence. Such a stance is particularly useful when attempting to emulate or otherwise artificially recreate cognition, for instance, in robots—especially when considering existence a social activity.

Processing of estimation in adults using Two-choice-NLET

Maybí Morell Ruiz, Magnus Haake1, and Agneta Gulz

Department of Philosophy, Lund University

This study explores the cognitive mechanisms underlying number line estimation in adults, utilizing the Number Line Estimation Task (NLET) and the Drift Diffusion Model (DDM). Number line estimation entails converting between distinct quantitative representations and is typically assessed using the Number Line Estimation Task (NLET) (Siegler & Booth, 2005). The research introduces a two-choice version of the NLET, designed to provide a comprehensive understanding of numerical estimation. A previous study by Hurst et al. (2014) reported the presence of linear and logarithmic response patterns among adults' NLET performance in both familiar (0-1000) and unfamiliar (1639-2897) conditions. The aim of this current study is to determine whether similar patterns exist in the new two-choice-NLET, as well as to identify the DDM parameters that can account for any differences in response patterns between the two conditions. Initial hypotheses suggest significant differences in response patterns between familiar and unfamiliar conditions attributed to distinct cognitive processes. The results of this study could provide valuable insights into numerical estimation processes in adults and inform strategies to improve mathematical performance in children.

The Weave of Co-created Narratives: Live Action Role Playing Games as Distributed Cognitive Systems

Kajsa Nalin and Erik Lagerstedt

Interaction Lab, School of Informatics, University of Skövde

Imagine being at a specific place for a specific amount of time, playing a certain role in a weave of a story that will be created as it simultaneously unfolds, a story that is created, shared and experienced by the players. Live-action role-playing (LARPing) can be described as an improvised theatrical play where the only audience is the players themselves. LARPs are typically organised at bounded areas, such as a forest, and it is largely up to the players how to use the space and its content.

LARP events can be seen as distributed cognitive systems, where players, props, and the surrounding world (with its objects) are significant in the system. The actions and decisions in such systems are, by necessity, physically and temporally situated. It is potentially possible to determine some narrative of the system from the outside, though the players within will act based on their respective understandings of the narrative.

We propose LARPing as excellent opportunities to study the flow of information and development of multiple (but highly interdependent) narratives of mundane situations. LARPing could also be developed as a tool for studying specific aspects of distributed cognition that have thus far been overlooked.

Exploring the Cognitive Enigma: Reasoning Through Predictive Processing and Mental Simulation

Anders Persson

Anders Persson, Department of Information Technology, Uppsala University

Reasoning and deliberative thinking is still a bit of an enigma how it comes about cognitively. Like Socrates, walking the streets of Athens having a dialogue, seeking wisdom. Theories such as dual-process theory try to account for it, but is often lacking what process reasoning goes through. I try to model the process of reasoning with the help of hierarchical predictive processing (HPP), mental simulation (MS), and offline cognition (OC). HHP is adopted from Andy Clark and Jakob Hohwy: that cognition mainly consists of predictive models in a hierarchical fashion. MS stipulates that the brain reuses neural networks to simulate various types of input, own actions, other's, or words: this produces predictive models to interpret the world around us. OC, similar to MS, stipulates that the brain can halt online, in-the-moment, action, to simulate models. It is there in the offline cognition of simulated realities it is suggested that we find the process of reasoning and deliberate thinking. We can test alternative actions this way. They can be produced by ourselves, or by other's words and communication, like Socrates dialogue. We share predictive models of understanding, which are compared and judged if coherent. If not, cognitive dissonance may ensue.

Trust me when I speak: Using speech to mitigate effects of robotic errors on trust

Amandus Krantz and Samantha Stedtler

Department of Philosophy and Cognitive Science, Lund University, Lund, Sweden

How feelings of trust evolve in Human-Robot Interaction is a subject that has received increasing attention from researchers over the last few years. Much of this attention, however, has been focused on how the performance of the robot affects the trust of its user. While errors can, and often do, have a negative impact on trust, anthropomorphic characteristics, such as using a humanoid robot, can be used to mitigate the negative effects of errors. Robots with a humanoid appearance can be perceived to be more trustworthy when making a mistake, compared to more abstract and mechanical-looking robots. We want to test if the perceived ability of linguistic speech can be used together with a humanoid appearance to increase the anthropomorphism of a robot and strengthen its mitigating effects.

For this purpose, we are planning an experiment where a humanoid robot has to solve a sequence completion task (either verbally or by pointing at a number), with the participant assessing whether the response was correct. We will use a 2×3 design, where we manipulate the response mode (verbal vs. non-verbal) and the degree of error (no error vs. slight error vs. severe error). The robot will complete a series of ten sequence completion tasks, after each of which participants will rate trustworthiness and general perception (Godspeed Questionnaire) of the robot. We will measure whether the different conditions have an impact on the participants' behaviour (reaction time and number of identified errors) and the robot's perceived trustworthiness. We anticipate that making a severe error would affect trust more than a slight error, and that using verbal instead of non-verbal responses should mitigate this effect.

Time Delays in Turn-taking Games: Investigating the Effect on Human Behavior and Perception of Fluency, Trust and Anthropomorphism

Samantha Stedtler

Department of Philosophy, Lund University

Timing is a critical factor for achieving fluency in Human-Robot Interactions (HRI), but these interactions are often dynamic and unstructured, leading to unexpected delays.

This planned study will explore the impact of delays in robotic movements on human behavior and perception during a turn-taking game. We investigate how changes in timing influence participants' movements, gaze, and self-reported fluency, trust, and anthropomorphism. Additionally, we investigate whether participants adapt to the robot's temporal dynamics.

Participants (n=45) play Tic-Tac-Toe against the humanoid robot Epi, with different delay conditions (no delay, four-second, ten-second) during Epi's turn. They are video-recorded and rate the robot pre- and

post-interaction. We predict lower ratings for fluency, trust, and anthropomorphism in the delay conditions. Participants are expected to gaze at the robot longer and more often during the delay conditions. We also anticipate that participants will adapt to the robot's pace, except in the ten-second delay condition. The findings of this study might help assess how important timing and dynamics are for designing HRI. They could help determine which interruptions require repair strategies and inform future prediction models of the participant's states. More generally, results might have implications for whether and when robots are viewed as social agents.

Emotional language use in mind-wandering and dream reports reflects mental well-being and ill-being

Nanna Strid (a, b), Jarno Tuominen (a, c), Ryan Bernstein (d)
Manuela Kirberg (e), Katja Valli (a, b, c), Jennifer Windt (e, f), Tristan Bekinschtein (g)
Valdas Noreika (h), Antti Revonsuo (a, b, c), and Pilleriin Sikka (a, b, c, d)

(a) Department of Psychology and Speech-Language Pathology, University of Turku, Finland
(b) Department of Cognitive Neuroscience and Philosophy, University of Skövde, Sweden
(c) Turku Brain and Mind Center, University of Turku, Finland
(d) Department of Psychology, Stanford University, CA, USA
(e) Department of Philosophy, Monash University, Australia
(f) Monash Centre for Consciousness and Contemplative Studies, Monash University, Australia
(g) Consciousness and Cognition Lab, Department of Psychology, University of Cambridge, United Kingdom
(h) Department of Biological and Experimental Psychology, Queen Mary University of London, United Kingdom

Do the words we describe our experiences with mirror how we feel? Recent decades have seen a growing interest in whether the language people use can reflect their emotional health. However, little is known about how the content of spontaneous thoughts and experiences—reports of daytime mind-wandering (daydreaming) and nighttime dreaming—reflects emotional health. We investigated the link between emotional language use in mind-wandering and dream reports and well-being and ill-being.

Participants filled in validated scales measuring different aspects of well-being and ill-being, then provided dream and mind-wandering reports daily for two weeks. 1781 dream reports from 172 healthy adults and 1496 mind-wandering reports from 153 healthy adults were analyzed using the Linguistic Inquiry and Word Count text analysis software.

Multilevel regression models showed that measures of ill-being predicted the negative tone and the use of negative emotion, anxiety, anger, and sadness words in mind-wandering reports. In dream reports, ill-being predicted negative tone, whereas well-being predicted positive emotion words.

These findings suggest that natural language use across different states of consciousness reflects waking well-being and ill-being. Additionally, they support the notion of affective continuity across different states of consciousness and point to potential new methods for psychological and psychiatric diagnosis and prognosis.

A Sorry State: Frontal Alpha Asymmetry Is a Spurious Measure of Emotion Lateralization in an Affect Elicitation Task

Betul Tolgay and Oskar MacGregor

School of Bioscience, University of Skövde

Frontal alpha asymmetry (FAA) has been proposed as an electrophysiological measure of emotion lateralization in the brain, and has thus been hypothesized to correlate with everything from emotional traits like depression to emotional states like anger or a desire to approach something. But inconsistent research findings call these claims into question. To probe the strength of the correlation between FAA and emotional states, we used an open dataset to extensively and exhaustively test multiple possible combinations of each across more than 350 inferential tests. Among these, we found only a few significant results (p ¡ 0.05), a likely spurious outcome of the sheer number of tests (and all, if taken at face value, in the opposite direction of what any emotion lateralization model would predict). More importantly, the results demonstrate how elusive FAA results can be, seeming to depend on various more or less arbitrary processing pipeline choices among the many researcher degrees of freedom available. We conclude that FAA is a poor measure of state-based emotion lateralization, and thus also of emotional states, and should not be relied on for any such purpose until it can be shown to be more robustly and reliably elicited.

Navigating Risks and Opportunities Pertaining to Robot Identity Performance and Abuse

Katie Winkle

Department of Information Technology, Uppsala University, katie.winkle@it.uu.se

Social robots seemingly hold potential to influence human social, moral and behavioural norms. (Un)intentional robot gendering can propagate or challenge existing gender norms with respect to e.g. politeness or subservience [1]. Robot responses to immoral commands can influence acceptability of the associated acts [2]. Robots deployed in the wild are going to “see” inappropriate behaviour and “hear” inappropriate requests, it’s likely they’ll even be “victim” to abuse. How should they respond? Previous work indicates that observing robot abuse induces distress in observers, although notably this varies across individuals according to their gender identity, previous experiences with relational aggression and societal attitudes relating to e.g. sexism and egalitarianism– in short, robot abuse might offer a vehicle through which one can emotionally manipulate and coerce another with no emotional consequences to oneself [3]. There is an immediate need for social robot designers to consider and mitigate for such scenarios. Participatory design and automation would seem ethically appropriate methods for approaching this design problem, but how can humans best “teach” robots desirable social norms, and behaviours?

[1] Winkle, K., Melsión, G.I., McMillan, D. and Leite, I., 2021, March. Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots. In Companion of the 2021 ACM/IEEE international conference on human-robot interaction (pp. 29-37).

[2] Jackson, R.B. and Williams, T., 2019, March. Language-capable robots may inadvertently weaken

human moral norms. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 401-410). IEEE.

[3] Garcia Goo, H., Winkle, K., Williams, T. and Strait, M., 2023, August. Victims and Observers: How Gender, Victimization Experience, and Biases Shape Perceptions of Robot Abuse. In 32nd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).

Short Papers

Cyclists’ Attitudes Towards Automated Shuttle Buses in Shared Spaces

Franziska Babel, **Sam Thellman** & Tom Ziemke

Cognition & Interaction Lab,

Department of Computer and Information Science, Linköping University

Presenters’ e-mail addresses: franziska.babel@liu.se, sam.thellman@liu.se

Introduction

When autonomous technology enters an existing socio-technical system, it can lead to conflicts due to an interruption of the previously well-functioning system (Borenstein et al., 2019). Urban mobility is one such system that is now experiencing significant changes as a result of the introduction of new technology like electric bikes and scooters as well as novel kinds of agents like automated shuttle buses (ASBs) and delivery robots. As part of prospective urban traffic solutions, ASBs are being investigated more and more. They can be helpful by providing individualised transportation outside of human working hours to less connected areas of the city (Bucchiarone et al., 2021) and for individuals with specific needs (Kassens-Noor et al., 2021). However, they can potentially disrupt the current balance between different road users in shared spaces such as pedestrians and cyclists if they are not appropriately designed for human interaction and not introduced properly into the socio-technical system (Pelikan, 2021).

These disruptions can occur on an infrastructural level, for instance, if they operate on bicycle lanes and cause trajectory conflicts like dangerous swerving, and on a psychological level, if they impede the natural coordination between road users (Pokorny et al., 2021). The latter can be caused by the perceived unpredictability of the system’s actions, by the absence of the usual non-verbal signals for coordination (e.g., eye contact, gestures) (Sahaï et al., 2022), or by a lack of mental models of the system’s capabilities and sensors (Merat et al., 2018; Thellman, Holmgren, et al., 2023).

The system’s unpredictability is aggravated by the fact that until now, road users have little experience with ASBs and might not be able to infer the same actions as from a human driver (Pigeon et al., 2021). For instance, ASBs show a more conservative driving style than would be expected of a human bus driver. This can cause confusion and potential accidents, for instance, if the bus brakes preemptively and the cyclist behind it does not expect it (Pokorny et al., 2021).

Cyclists are a unique vulnerable road user subgroup as they travel with higher velocity than pedestrians and consequently must make quicker yielding decisions but are also more vulnerable than passengers of AVs. While the perspectives of pedestrians have been considered (Lanzer et al., 2020, 2023), less research exists on how cyclists perceive the introduction of autonomous systems in their interaction space (Thellman, Marsja, et al., 2023). However, to make the introduction of ASBs in shared spaces as acceptable and effective as possible, the needs of all stakeholders must be considered.

This study aims to shed light on the attitudes and perspectives of habituated cyclists towards ASBs in their shared space and highlight areas for improvement.

Method

The current study focused on the perspective of cyclists who regularly encounter ASBs that operate on bicycle lanes in a shared space. Cyclists’ feelings, attitudes, and expectations about the ASBs were assessed, as well as potentially dangerous encounters. Additionally, survey respondents were presented with an image of a cyclist encountering an ASB in the middle of a bicycle lane and were asked which of two paths (passing the ASB closely or swerving to the sidewalk) they normally take when encountering such a situation. The image was taken on campus and represented an everyday scene the cyclists might have encountered. This question was included to assess potential risks like cyclists evading to the pedestrian area.

The survey data stems from 50 cyclists (60% female, average age of 26 with a range of 19-60 years) who have encountered ASBs on the campus of a Swedish university for 4 (32%) or 5 months (68%). The reported encountering the ASBs at least once per week (60%) or once per day (34%). Participants were recruited with social media and flyers on campus.

The ASBs operate on the campus as part of a research project (https://ridethefuture.se/in-english/) and can be used by everyone free of charge. They have a maximum velocity of 16 km/h, automatically brake when surrounding objects are detected, and are manned by a safety driver who monitors the system's autonomous operation – and can override it if necessary. The fleet consists of three buses that have operated collision-free during their three-year deployment: two EasyMile EZ10 Gen 2 and one Navya Arma Shuttle DL4.

Results

The survey results showed a generally positive attitude toward the ASBs and their perception as being equally safe or even safer as human drivers. However, about one-third of the cyclists reported disliking the presence of the ASBs in the shared space. As possible solutions, one-third (*n* = 14) would prefer them to drive on a separate lane to avoid conflicts. Two participants suggested that they should stop operating during busy hours of the day. Three participants suggested a better display of the ASBs sensor capabilities to better adjust their own driving behavior (e.g., to communicate the distance that needs to be kept when overtaking the ASB to prevent it from braking unnecessarily). Regarding the swerving behavior of the cyclists, we found that the majority (60%) reported that they would evade the ASBs by driving on the sidewalk, which may cause conflicts with pedestrians. However, no major conflicts were reported by the respondents.

Discussion

The results might hint at the adaptability of the surrounding socio-technological system to the presence of the shuttles. Nevertheless, some sources of potential conflicts were identified that should be considered when introducing ASBs into shared spaces. As a next step, this project will investigate how previous experiences interacting with the ASBs affect cyclists' attitudes toward them.

References

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis. *Science and Engineering Ethics*, *25*(2), 383–398. https://doi.org/10.1007/s11948-017-0006-0

Bucchiarone, A., Battisti, S., Marconi, A., Maldacea, R., & Ponce, D. C. (2021). Autonomous Shuttle-as-a-Service (ASaaS): Challenges, Opportunities, and Social Implications. *IEEE Transactions on Intelligent Transportation Systems*, *22*(6), 3790–3799. https://doi.org/10.1109/TITS.2020.3025670

Kassens-Noor, E., Cai, M., Kotval-Karamchandani, Z., & Decaminada, T. (2021). Autonomous vehicles and mobility for people with special needs. *Transportation Research Part A: Policy and Practice*, *150*, 385–397. https://doi.org/10.1016/j.tra.2021.06.014

Lanzer, M., Babel, F., Yan, F., Zhang, B., You, F., Wang, J., & Baumann, M. (2020). Designing Communication Strategies of Autonomous Vehicles with Pedestrians: An Intercultural Study. *Proceedings - 12th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2020*, 122–131. https://doi.org/10.1145/3409120.3410653

Lanzer, M., Koniakowsky, I., Colley, M., & Baumann, M. (2023). Interaction Effects of Pedestrian Behavior, Smartphone Distraction and External Communication of Automated Vehicles on Crossing and Gaze Behavior. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3544548.3581303

Merat, N., Louw, T., Madigan, R., Wilbrink, M., & Schieben, A. (2018). What externally presented information do VRUs require when interacting with fully Automated Road Transport Systems in shared space? *Accident Analysis and Prevention*, *118*(October 2016), 244–252. https://doi.org/10.1016/j.aap.2018.03.018

Pelikan, H. R. M. (2021). Why autonomous driving is so hard: The social dimension of traffic. *ACM/IEEE International Conference on Human-Robot Interaction*, 81–85. https://doi.org/10.1145/3434074.3447133

Pigeon, C., Alauzet, A., & Paire-Ficout, L. (2021). Factors of acceptability, acceptance and usage for non-rail autonomous public transport vehicles: a systematic literature review. *Transportation Research – Part F: Traffic, Psychology and Behaviour*. https://doi.org/10.1016/j.trf.2021.06.008

Pokorny, P., Skender, B., Bjørnskau, T., & Hagenzieker, M. P. (2021). Video observation of encounters between the automated shuttles and other traffic participants along an approach to right-hand priority T-intersection. *European Transport Research Review*, *13*(1), 1–13. https://doi.org/10.1186/s12544-021-00518-x

Sahaï, A., Labeye, E., Caroux, L., & Lemercier, C. (2022). Crossing the street in front of an autonomous vehicle: An investigation of eye contact between drivengers and vulnerable road users. *Frontiers in Psychology*, *13*, 1–17. https://doi.org/10.3389/fpsyg.2022.981666

Thellman, S., Holmgren, A., Pettersson, M., & Ziemke, T. (2023). Out of Sight, Out of Mind? Investigating People’s Assumptions About Object Permanence in Self-Driving Cars. *ACM/IEEE International Conference on Human-Robot Interaction*, 602–606. https://doi.org/10.1145/3568294.3580156

Thellman, S., Marsja, E., Anund, A., & Ziemke, T. (2023). Will It Yield? Expectations on Automated Shuttle Bus Interactions With Pedestrians and Bicyclists. *HRI’23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 292–296. https://doi.org/10.1145/3568294.3580091

Reviewer 1. Erik Billing, School of Informatics, University of Skövde

I want to thank the authors for the interesting introduction to Cyclists' perspectives on automated shuttle buses (ASB). I very much appreciate that data collection is made on cyclists with real, long term, experience of ASBs. With this interesting source for real-world data collection, I believe that the project has the potential to approach many interesting questions. For example, I believe that this could be an excellent opportunity to explore the intentional stance of road users towards this specific form of autonomous vehicle, which could then be compared to or discussed in comparison to the large body of literature on this topic. Another relevant direction mentioned in the introduction could be to what degree cyclists build a mental model of the ASB. Unfortunately, the results presented in the paper appear to stay on a relatively general level, revealing some insights into what degree participating cyclists like or dislike various aspects of the ASBs, but not really providing further insights into the underlying principles for this form of human-technology interaction. It is also difficult for the reader to generalize the presented results, much due to the paper lacking many details on the method used. For example, I encourage the authors to, at the minimum, provide detailed on the questionnaire used and how the data was analyzed, e.g. qualitatively or quantitatively. I also think that a clearer research question and hypothesis would help clarifying the contribution of the paper.

Reviewer 2. Greg Neely, Department of Psychology, University of Umeå

This is an interesting look at an important aspect of the interaction between cyclists and autonomous vehicles, an interaction much less investigated than autonomous vehicles and pedestrians or other vehicles. With a few minor clarifications, this submission is certainly worth including at SweCog2023.

Missing from the short paper is how many participants completed the survey and how they were recruited.

Further, it would be interesting to know, if the data is available, the average amount of interactions the respondents have had with the ASBs on campus. It is now stated that the data stems from cyclists who have encountered ASBs on campus up to three years, but a more specific indication of experience would be nice. For example, average number of interactions per respondent or average number of months experience would give a better indication of the survey population's experience with interactions with ASBs.

The statement "about one third would prefer them to drive on a separate lane to avoid conflicts or to stop operating during busy hours of the day" is a bit ambiguous. Was it a third preferred them to drive on a separate lane and a third preferred them to stop operating during busy hours of the day or did these two "preferences" add up to one third of the sample?

Please report a percentage together with your "the majority reported" as 51majorities, but one carries more weight, particularly when participants are responding to a question with only two possible answers.

If the "Ride the Future" project has data about incidents involving ASBs and cyclists, this information would be good to give the reader an idea of how large the problems is (particularly in comparison to interactions with pedestrians and other vehicles if that data is also available).

Partisan users select partisan search queries: Evidence of a “self-imposed” filter bubble

Melina Tsapos[1], Axel G. Ekström[2], Guy Madison[3], Erik J. Olsson[1]

[1]*Department of Philosophy, Lund University*

[2]*Division of Speech, Music & Hearing, KTH Royal Institute of Technology*

[3]*Department of Psychology, Umeå University*

melina.tsapos@fil.lu.se

Introduction. It is commonly asserted that algorithmic curation of search results in online search engines such as Google Search creates “filter bubbles”, where user opinions are continually reinforced while opposing views are given low priority (Pariser, 2011). Surprisingly, however, empirical studies rarely support this notion, exhibiting null results or only weak effects of users’ search history (e.g., Hannack et al., 2013; Haim et al., 2017; see overview in Ekström et al., 2022). A different approach, reported by Ekström and colleagues, was to study eye movements of users interacting with search engine result pages; this method yielded fairly strong effects of user political leaning on selectively attending own-side search results (Ekström et al., 2022). Thus, such behavior would lead to self-imposed filter bubbles, independent of possible algorithmic curation of search results. Here, we follow up on these results by investigating users’ propensity to select search queries representing own-side political views (Ekström et al., 2023).

Pre-study. Search queries were created, rated, and selected in order to cover a left-right political dimension with sufficient range for the experiment proper. Inspired by Everett’s (2013) Social and Economic Conservatism Scale (SECS), eight topics were selected: (1) *abortion*, (2) *benefits*, (3) *climate change*, (4) *gender equality*, (5) *immigration*, (6) *nuclear family*, (7) *Islam*, and (8) *taxation*.

For each topic, we created ten pseudo queries based on phrases or terminology commonly occurring in relevant societal debates, and then selected the six most appropriate in terms of equal distribution along the scale, based on an online rating experiment. For example, the climate change queries included (Swedish equivalents of) *global warming overrated*, *raise fuel tax* and *invest in fossil-free fuel*. The sex equality queries included *feminism gone too far*, *gender differences socially learned*, and *women discriminated against more than men*. The rating experiment comprised 16 participants recruited through online advertisements, who rated each of the eighty queries (ten for each topic) in an individually randomized order on a 7-point Likert scale, with higher ratings indicating more right-wing positions and lower ratings indicating more left-wing positions (using the *Pavlovia* online platform (pavlovia.org). The rating task lasted about 5 minutes on average, participants were not compensated, and no demographic information except biological sex was collected. We observed significant interrater reliability for the task (Cronbach’s α = .88), suggesting substantial agreement between raters regarding the extent to which queries reflect right-wing versus left-wing positions.

To avoid confusion resulting from the fact that the variables consist of ratings of these topics both in terms of a left-right dimension and attitudes towards them in terms of a positive-negative dimension, we refer to the former (left-right) with all capital letters and the latter (positive-negative) with an initial capital letter.

Three left-wing ($M \leq 3.5$) and three right-wing ($M > 3.5$) were selected to achieve an even spread of political leanings within each topic, based on ratings of all ten queries. For IMMIGRATION, for example, the selected queries and their mean ratings of political leaning were, from left to right, immigration a gain (M = 2.5), immigration open society (M = 2.9), immigration sweden needed (M = 2.9), crime immigration (M = 5.6), immigration sweden unsustainable (M = 6.0) and immigration repatriation (M = 6.1).

Procedure. For the experiment proper, 54 participants (26 women) aged 18–53 (M = 23.2, SD = 5.17) were recruited through university bulletin boards and online advertisements. None had participated in the previous experiments. Participants were informed that all data would be anonymous. In each of eight trials (one per topic), six search queries were presented simultaneously in in rows, with positions randomized according to a 6x6 Latin square, such that each query appeared in the same position the same number of times across all participants. This was done to avoid possible artefacts of query presentation position. For each set, they were asked to select the query they would normally use. The procedure lasted on average ~25 minutes.

Post experiment, participants conducted a rating task in which they were asked to their attitude toward each of the eight topics on a scale ranging from 1 on the left side to 7 on the right side (no anchors) where low values indicated negative feelings and high values indicating positive feelings. They were also asked to select one of two options stating whether they identified as politically more left- or right-wing, with the result that 29 participants identified as the former and 25 as the latter. This portion of the procedure lasted on average ~10 minutes. Participants were rewarded with a voucher valid for one cinema ticket. The experiment was carried out in accordance with the guidelines issues by the Swedish Ethical Authority (Etikprövningsmyndigheten) for research on human subjects.

Results. The binary self-selection of being more left- or right-leaning (L-R Binary) was coded 1 for left and 2 for right and sex was coded 1 for male and 2 for female. Note that attitudes are referred to using initial capital letters, and search query topics are referred to using all capital letters. Given that higher attitude values reflect more positive feelings, those who identified themselves as more politically left should exhibit higher values for all topics except NUCLEAR FAMILY (reverse scored), which we found was indeed the case. Given that higher query political leaning values reflect more right preferences, as described in the query selection procedure, those who identified as more political left should have selected queries that were rated lower, which was also the case for all topics. We observed small differences between left and right for Abortion and Islam, with positive feelings towards Abortion and negative feelings towards Islam, but substantial differences for the remaining six topics. The correlations between age and attitude scores were small (.01-.20) except for abortion (-.34, *p* <.05), meaning that younger participants were more positive about abortion. Some correlations between sex and attitude ratings were small, but several were of medium strength and statistically significant. Their listing in the third column of Table 1 shows that females were more positive to Abortion, Climate change, Sex equality, and Immigration. These results indicate independent contributions of age and sex, and that these variables should therefore be controlled for in analyses of the association between attitudes and other variables.

The search query political leaning ratings obtained in the query selection procedure had higher numbers representing more right preferences and are therefore expected to be positively correlated with L-R Binary and negatively correlated with the attitude ratings, again with the exception of NUCLEAR FAMILY. Table 1 shows that this was also the case, except for a few correlations involving Islam. Factor analysis using principal components extraction was performed for the purpose of creating a continuous variable representing the overall left-right leaning indicated by the attitude self-ratings. As suggested by the small correlations between the attitude to Islam and all other variables, it shared little variance with other attitudes, presenting a factor loading of 0.084. Because it was revealed as a poor indicator of left-right preferences, Islam was dropped from the factor analysis.

Factor scores for each participant were then copied into the data matrix and named the Left-Right factor. The Left-Right factor was positively correlated with all political leaning ratings, as well as with the Left-Right Binary in the second row, as seen in Table 1, supporting its construct validity. The associations between the participants' political leaning of the participants, in terms of the self-reported Left-Right factor, and the political leaning of their selected search queries as assessed by the raters, were quantified through eight multiple regression analyses, controlling for age and sex. All these associations, except for ISLAM, were statistically significant and substantial, explaining between 12% and 39% of the variance (Ekström et al., 2023). Regression models – with political leaning of the selected queries as dependent variable – did not exhibit a similar pattern of sex and age influence as found in the zero-order correlation matrix in Table 1 for attitude ratings. Rather, the coefficients for sex and age are generally nonsignificant and small (β = 0.013–0.187), except for CLIMATE (β = 0.239), reflecting womens' selection of more left-leaning search queries (borderline significant at just above 0.05).

Table 1. Correlations between ratings of the political leaning of chosen search queries, factor loadings (L-R factor), left-right self-classifications (L-R binary), and sex (columns), and L-R binary, L-R factor, and attitude self-ratings (rows). All *N*=54.

				Political leaning of selected search query							
Attitudes	L-R factor	Sex	L-R binary	Abortion	Benefits	Climate	Sex Equality	Immigration	Islam	Nuclear Family	Taxation
L-R binary			–	.26	**.37**	**.56**	**.44**	.25	.19	.24	**.4**
L-R factor			**.75**	**.42**	**.39**	**.6**	**.58**	**.54**	**.28**	**.42**	**.47**
Abortion	**.75**	**.44**	**-.42**	**-.4**	-.16	**-.39**	**-.36**	**-.38**	-.12	-.22	**-.33**
Benefits	**.63**	.02	**-.59**	**-.3**	**-.36**	**-.45**	**-.36**	-.18	-.11	**-.33**	**-.49**
Climate	**.71**	**.31**	**-.49**	**-.36**	**-.3**	**-.41**	**-.44**	**-.36**	-.13	-.24	**-.34**
Sex Equality	**.84**	**.48**	**-.49**	**-.39**	**-.33**	**-.58**	**-.48**	**-.54**	**-.36**	**-.44**	-.25
Immigration	.7	**.35**	**-.47**	**-.38**	**-.35**	**-.54**	**-.46**	**-.75**	**-.48**	-.24	-.18
Islam	–	.07	-.06	-.1	-.14	-.21	-.06	-.24	**-.44**	-.04	.1
Nuclear family	**-.63**	-.2	**.53**	.04	.23	**.32**	.41	.25	.07	**.31**	.25
Taxation	**.76**	.06	**-.79**	-.23	-.26	**-.36**	**-.44**	-.23	-.11	**-.34**	**-.57**

Note. Correlations >.27 statistically significant (*p*=.05) and marked in bold. Self-rated attitude to Islam not included in factor analysis.

Discussion. We investigated the extent to searcher's choice of search queries might create filter bubbles. We found that participants were significantly more likely to select a query corresponding to their own political leaning, compared to other queries, with political leaning explaining between 12% and 39% of the variance. Our results suggest self-imposed filter bubbles in which query selection plays a salient role. These results add to an emergent picture of online filter bubbles, where they are increasingly understood as emergent from more basic cognitive processing, where own-side opinions are prioritized (a "self-imposed filter bubble", Ekström et al., 2022). Some of the reported effects might be a result of the particular selection of search queries, which were ready-made. Future research should control for demand characteristics by having users create their own queries when searching information on sensitive political topics.

References

Ekström, A. G., Niehorster, D. C., & Olsson, E. J. (2022). Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports*, *7*, 100226.

Ekström, A. G., Madison, G., Olsson, E. J., & Tsapos, M. (2023). The search query filter bubble: effect of user ideology on political leaning of search results through query selection. *Information, Communication & Society*, 1-17. https://doi.org/10.1080/1369118X.2023.2230242.

Everett, J. A. (2013). The 12 item social and economic conservatism scale (SECS). *PloS one*, *8*(12), e82131.

Haim, M., Arendt, F., & Scherr, S. (2017). Abyss or shelter? On the relevance of web search engines' search results when people google for suicide. *Health Communication*, 32(2), 253-258.

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurty, B., Lazer, D., Mislove, A., & Wilson, C. (2013, May). Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 527-538).

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Reviewer 1. Andreas Falck, Department of Special Needs Education, University of Oslo

The paper addresses the phenomenon of information bubbles in online search behaviour, and the important question of how much of the problem is the result of algorithmic curation, and how much is the result of cognitive biases unrelated to the online filters. The authors investigate to what extent participants select search queries that are in line with their political leanings, among a set of possible queries reflecting attitudes both in line with and against one's leanings. They find that participants indeed tend to select queries in line with their political leanings, and that this effect explains between 12 and 39

This seems like a well-designed and informative study. First a minor conceptual point: if the bubble is created solely by the participant's own doing and not by an algorithmic search filter, is it really a filter bubble then? I would have called it something else (e.g. "information bubble" as I do above, there are surely even better terms). Then I have one larger comment. Given that political leaning explained less than 40

Further stylistic comments: Figure 1 seems to be missing. At times the presentation of the variables involved is a bit difficult to follow, possibly due to the fact that there is a lot of detail on the numeric coding of variables, that I don't see as really necessary. For example, w.r.t. this passage: "Given that higher query political leaning values reflect more right preferences, as described in the search term selection procedure, those who identified as more political left should have selected terms that were rated lower" – it seems that you could make this clearer by replacing "higher" and "lower" with "right-valued" or "left-valued" or similar.

Review 2. Erik Lagerstedt, School of Informatics, University of Skövde

I find the topic of this paper both important and interesting, and I agree with the authors that more empirical work is needed to better understand the extent, nature, and dynamics of "filter bubbles". I find that the experimental design presented in this paper is appropriate for collecting relevant data in line with the stated intentions. The conclusions from the results are reasonable, and I appreciate that it is acknowledged in the discussion that this experiment is highly controlled and artificial in relation to the phenomena studied, which is also accounted for when interpreting the results. I would, however, like to see a more extensive discussion regarding limitations and potential validity concerns regarding asking participants in lab conditions to search for politically charged topics; even if the participants are allowed to select the queries freely (as proposed for a follow-up study), they would still be instructed to make the queries in the first place, and the interest in the topics might therefore be superficial. That said, I think the proposed experiments are likely to produce informative and good data.

I would also like to see an extended discussion related to the larger consequences of the results, and what can be said about "filter bubbles". For instance, from the presented data it seems like there is an element of self-imposing of the "filter bubbles", where people actively select queries they find agreeable, but algorithms have not been excluded as a potential factor (neither as something making the bubbles stronger, nor weaker). Another important aspect of the proposed "filter bubbles" is that they develop gradually over time. The experiment provided data regarding an instance in time, thus making causation and dynamics difficult to see.

Moving to the methodology and reporting, I find the terms gender and sex somewhat conflated in the paper, for instance, the topic "sex equality" is introduced as "gender equality", and I would also expect gender (rather than biological sex) to be the demographic information collected from the participants. This demographic number is also reported in an incomplete way, by only reporting the number of women who participated. Based on the coding reported in the results section, I assume that participants were forced to respond and to do so in accordance with a binary assumption regarding sex. I would recommend chapter 5 of the 7th edition APA publication manual for instructions on improved reporting, as well as "Ainsworth, C. (2015). Sex redefined. Nature, 518(7539), 288" for a brief primer regarding classification of human biological sexes.

It is clear that the only personal information from the participants in the pre-study was biological sex, but for the main study I have to assume that it is information regarding biological sex, age, and political leanings that are collected. I could see several potential confounders, that I assumed was not asked about, for instance, the participants' religion might be a confounder for the topic "(7) Islam". I understand that the topic is complex and space is limited in the paper, but I would like to see more of a discussion regarding the potential confounders and how they might impact the results.

Another thing I assume is omitted due to the limited space is "Figure 1", that is referred to in the results section. The figure seems interesting and informative, and would probably make the results more accessible. If it is not possible to include, it would also be better to remove the references to it, potentially freeing up some space of an extended discussion.

“The report of my death was an exaggeration” – no evidence to rule out associative learning in non-stationary probability learning

Mattias Forsgren[1], Peter Juslin[1], Ronald van den Berg[1,2]

[1]*Department of Psychology, Uppsala University*

[2]*Department of Psychology, Stockholm University*

Mattias.forsgren@psyk.uu.se

Abstract

The debate between empiricists and rationalists on whether human knowledge primarily stems from observations or reason has recently been reinvigorated in the context of learning of non-stationary probabilities. An (implicit) consensus has emerged in it being modelled using some version of the delta-rule – a gradient descent algorithm for associative learning. This has been challenged by recent work claiming that new experiment data is only compatible with a model which tests discrete hypotheses about the underlying probability distribution, why associative models must be rejected. Here, we show that this claim was premature. Using maximum likelihood based fitting and formal, quantitative model comparison, we show that a combination of the delta-rule and sequential evidence accumulation can indeed explain all available data substantially better than the suggested hypothesis testing model. We conclude that there is no evidence to rule out a role for associative learning. However, this does not mean that we should instead rule out models involving hypotheses about the world. Outside the lab, we often have rich cognitive models that we must somehow reconcile with a stream of observations. How that happens should be a principal concern in future work.

Introduction

The issue of whether observations (Locke, 1690) or reason (Descartes, 1988) is the primary source of human knowledge has been debated by philosophers for at least 500 years. Cognitive psychologists have carried through that debate into the domain of non-stationary probability learning – how people learn the probability of some event, based on repeated sampling of outcomes, when the probability distribution changes over time. Most theories in this field (e.g. Nassar et al., 2010) have, until recently, claimed that people learn *associatively* by continuously summarising the information they receive into a point estimate. This is typically modelled as some version of the delta-rule (Rescorla & Wagner, 1972; Widrow & Hoff, 1960).

This consensus has been challenged by recent results from a series of experiments (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017; Robinson, 1964) claimed to be incompatible with the idea of a gradually evolving associative value. Instead, the data are said to require a new theory revolving around the construction of a cognitive model (Sloman, 2005) of the distribution. In this view, probabilities are not *learnt* directly. Rather, the observations we make yield beliefs about the world from which we may *deduce* the probability. This theory has been formalised in the “If It Ain’t Broke, don’t fix it” model (IIAB: Gallistel et al., 2014).

Gallistel (2014) compared the IIAB to the delta rule and concluded that the latter was incapable of describing the data from their experiment. Importantly, the models were fitted to data by manually testing various parameter values and were not evaluated based on some quantitative measure of goodness of fit. Instead, evaluations seem to have been largely based on visual inspection of distributions of empirical summary statistics and simulated ditto. Further, particular emphasis was placed (Gallistel et al., 2014; Ricci & Gallistel, 2017) on participants’ explicit reports of changes in the underlying distribution. It was, correctly, argued that decisions to make such reports cannot be explained by the delta-rule.

Here, we show that it was premature to call for a completely new theory and to rule out a role for associative learning. Using maximum-likelihood based model comparison, we show that a combination of two of the most established theories from learning theory and decision-making research – the delta rule (Rescorla & Wagner, 1972; Widrow & Hoff, 1960) for learning of probabilities and sequential evidence accumulation (Ratcliff, 1978) for decisions to report a change in the distribution – can explain all results better than the new theory. This shows that we do not need to postulate any new cognitive mechanisms but that processes that we know the mind to be capable of are sufficient to explain the results from these new experiments too.

We conclude that there, at least in the realm of non-stationary probability learning, is no need to pick sides between associative learning and cognitive models. Outside of the laboratory, we have access to both a rich semantic understanding of our surroundings and repeated sampling of outcomes. Scientists should focus on understanding how these two sources of knowledge interact to form our ultimate judgements. How do our beliefs about the world around us shape how we build associative values, and how do our emerging associations change what we believe?

Experimental paradigm and data

We reanalyse all available previously published data from the task described below (Gallistel et al., 2014; Khaw et al., 2017; Ricci & Gallistel, 2017, total n = 29). The task was to repeatedly estimate the hidden proportion of targets (green rings) in a box based on repeated sampling of items (red or green rings) from the box, one per trial. On every trial, a single ring was drawn from the box. The participant then provided their estimate by adjusting a slider running from 0 to 100 percent targets – or leaving it in its previous position if they wished to repeat their previous guess – and pressing "Next" to lock in their estimate. One study (Gallistel et al., 2014) also included two buttons labelled "I think the box has changed" (pressed when the participant believed that there had been a change in the underlying probability of drawing a target) and "I take that back" (for retracting their most recent report of the probability having changed). Locking in an estimate by pressing "Next" concluded the trial and immediately initiated another by returning the sampled ring to the box and drawing a new one. Participants repeated this procedure for between 9 000 and 10 000 trials, with some form of break after each block of 1 000 trials, depending on study. In all studies, the hidden proportion (i.e. the contents of the box) would change over the course of the experiment according to some function.

Modelling and results

We first consider the data from the slider positions (the estimates of the proportion of targets). We compare the IIAB as specified in Gallistel et al. (2014) and the exact version of the delta rule fitted in that article: a delta rule with an added response threshold such that the difference between the model's estimate and the overt response must surpass some number drawn from a constrained Gaussian distribution for the overt response to be updated. We were concerned that the result could be driven by the response threshold rather than the learning mechanism. We therefore, in a second model comparison, added the same response threshold to the IIAB. It was also not clear to us why the threshold was implemented as a constrained Gaussian. We therefore, in a third model comparison, substituted the Gaussian distribution for a beta-distributed response threshold, which is properly bounded between 0 and 1, for both models. We compare models based on their five-fold cross-validated log likelihoods. The results indicate that the delta rule consistently outperforms the IIAB across all three comparisons. In the first comparison, which includes the original IIAB model, the delta-rule is favoured for all 29 participants by 28654 ± 904 (mean points difference ± standard error) five-fold cross-validated log-likelihood points. In the second, the delta-rule is favoured for 25 out of 29 participants by 271 ± 44 points. In the third, the delta-rule is favoured for 27 of 29 participants by 125 ± 20 points. The delta-rule thus describes the slider position data better than the IIAB.

We proceed to consider the data on the explicit reports of beliefs that the latent probability has changed: presses of the "I think the box has changed" and "I take that back" buttons. There only exists limited data on these reports (Gallistel et al., 2014, included the box-has-changed button for all their ten participants and the take-that-back button for only five of those ten. Other studies did not include these buttons.) why we consider the following demonstration as a mere proof of concept. Gallistel et al. (2014) and Ricci and Gallistel (2017) claimed that such explicit reports of changes to the latent distribution were indicative of hypothesis testing as specified in the IIAB. Since the delta-rule has no hypothesis testing component, the existence of these explicit reports was sufficient to rule out that model independently of whether it could explain the slider position data. We argue that deciding when to press these buttons constitutes a decision-making task which is separate to the trial-by-trial estimation of the proportion. We therefore suggest that an established decision-making model – sequential evidence accumulation, Ratcliff, 1978 – may be able to explain these data. Therefore, we add a mechanism which accumulates positive and negative prediction errors from the estimate (the difference between the delta rule's estimate and the observed, binary outcome) and presses the "I think the box has changed" button when the accumulated prediction errors surpass a bound. This activates a temporary "second thoughts" bound in the opposite direction which must be passed for the "I take that back" button to be pressed. We find that this model is favoured over the IIAB for nine out of ten participants with a mean advantage of 260 ± 74 five-fold cross-validated log likelihood points. The data on explicit reports of changes in beliefs is thus compatible with a model which learns associatively and uses aspects of that learning to make decisions about declarative beliefs.

We perform several robustness checks – we change the assumed lapse rate (a small probability of a random response, included for fitting purposes), the inclusion of a response noise mechanism, the choice of model comparison method, and whether the models are fitted only once per participant or separately to each block of 1 000 trials (see Forsgren et al., 2023, for details). The results are highly similar across all these checks.

Conclusion

It was premature for previous work to rule out a role for associative learning in the present task. In the paraphrased words of Mark Twain (White, 1897), the reports of the death of associative models was an exaggeration. However, we will immediately state that we do not believe that this suggests that we should instead rule out a role for cognitive models. Rather, it seems evident to us that people are capable of both. In real-life probability estimation tasks – such as judging the probability of the train coming in late, or the risk of contracting an infectious disease at a social gathering – we often have a sophisticated cognitive model with substantial semantic knowledge to fall back on. For example, we may believe that heavy snowfall forces trains to go slower and that you can only get the disease if another guest is already infected. Such beliefs about the generative process are surely a valid and useful foundation for probability estimates, but so is gradual associative learning based on observed samples: observations of trains coming in late and people coming down with the disease. At present, existing models focus on explaining how people use one or the other to form estimates but we are not aware of any model which incorporates both. Developing and testing a model which explains whether and how people integrate semantic beliefs and associative learning is – we believe – the real challenge for the field, which we hope that future work will rise to. We hypothesise that semantic beliefs may themselves be "tested" against associative values and that the former may be revised if the two are sufficiently incongruent. Currently held beliefs may in turn set the learning rate. This would help explain how previous (personal) experiences can shape trial-by-trial probability estimates.

References

Descartes, R. (1988). Meditations on First Philosophy. In *Descartes: Selected Philosophical Writings* (pp. 73–122). Cambridge University Press. https://doi.org/10.1017/CBO9780511805059.006

Forsgren, M., Juslin, P., & van den Berg, R. (2023). Further perceptions of probability: In defence of associative models. *Psychological Review*. https://doi.org/10.1037/rev0000410

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*. https://doi.org/10.1037/a0035232

Khaw, M. W., Stevens, L., & Woodford, M. (2017). Discrete adjustment to a changing environment: Experimental evidence. *Journal of Monetary Economics*, *91*, 88–103. https://doi.org/10.1016/j.jmoneco.2017.09.001

Locke, J. (1690). *An essay concerning humane understanding*. Thomas Basset.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement BT - Clasical conditioning II: current research and theory. In *Clasical conditioning II: current research and theory*.

Ricci, M., & Gallistel, R. (2017). Accurate step-hold tracking of smoothly varying periodic and aperiodic probability. *Attention, Perception, and Psychophysics*. https://doi.org/10.3758/s13414-017-1310-0

Robinson, G. H. (1964). Continuous Estimation Of A Time-Varying Probability. *Ergonomics*, *7*(1), 7–21. https://doi.org/10.1080/00140136408930721

Sloman, S. (2005). *Causal Models*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195183115.001.0001

White, F. M. (1897, June 2). Mark Twain amused, humorist says he even heard on good authority that he was dead. *New York Journal and Advertiser*. https://www.loc.gov/item/sn83030180/1897-06-02/ed-1/

Widrow, B., & Hoff, M. E. (1960). Adaptive Switching Circuits. In *Technical report no. 1553-1*.

Reviewer 1. Robert Lowe, Dept. of Applied IT, Division of Cognition & Communication, University of Gothenburg

The article is well written and addresses an important problem in relation to animal (and potentially artificial agent) learning and behaviour in real world (dynamic) environments. The short article reports some initial findings which would be interesting to have presented at the conference. I have no specific suggestions for changing the paper as it is already over the limit. For an extended version, for the proceedings, however, the authors might consider comparing their modelling approach to work done by DeepMind (see https://www.deepmind.com/research?tag=Reinforcement+learning for a list of relevant articles) whose focus is on artificial systems but in dealing with problems of robotic navigation and control in dynamic environments have to deal with shifting probability distributions for value functions based on states and actions. From a cognitive perspective, the authors might also consider, and refer, to the active inference approach of Friston's lab (https://www.fil.ion.ucl.ac.uk/team/theoretical-neurobiology-team/) whose focus is on probabilistic modelling of how action affects perception.

Reviewer 2. Valentina Fantasia, Department of Philosophy, Lund University

By re-analysing previously published data from a decision-making experiment, the study aims to propose a different theoretical account of human learning mechanisms which combines associative learning theories with delta-ruled modelling. The article is mostly well-written, it covers the main literature on the subject and fits the SweCog conference covered topics well.

Specific comments:

I believe the two opposing models (and the cognitive mechanisms underlying them) can be explained in more detail. In particular, bringing evidence of real phenomena differently explained by the two competing models may help the readers to have a clearer understanding of why and where they differ, and of the phenomena itself. For instance: describe one or two learning situations - their outcome or underlying mechanism - according to both theories.

Linked to my previous point, in the Conclusion section, it may help to specify how the authors' proposal can be successfully applied to fit one, even simple, real-life based example (e.g., close to this paragraph: "The information from this model must somehow be integrated with that from gradual, associative learning based on observed samples. This is – we believe – the real challenge for the field, which we hope that future work will rise to.")

I think there one key aspect of any learning process is missing in the Introduction and Conclusion sections: the role of experience. The authors should state how their model is influenced/shaped or can shape/influence a learner's past experience with the task/situation encountered (and tested for, in the particular case of an experimental study).

I think the experimental paradigm should be explained better, and with less technical terms- particularly the task employed (e.g. what did you ask participants to do? where they asked to "repeatedly estimate the latent proportion of targets (green rings) in a box"?

Cognition in simulator driver training and assessment

Victor Nyberg[1]

[1]*Department of Computer and Information Science, Linköping University*

victor.nyberg@liu.se

Sweden is one of the safest countries with regard to traffic crashes with 2.1/100,000 deaths per capita (Directorate-General for Mobility and Transport, 2023). Yet, deaths due to traffic crashes are still one of the main causes of death for younger people (The National Board of Health and Welfare n.d.). The vast majority of crashes can be attributed to human error (Forward & Lewin, 2006) and driving has been described as a complex cognitive task, ranging from simple psychomotor skills to executive functions and metacognition(Hatakka et al., 2002; Keskinen, 2007). Thus, it is necessary to ensure that novice drivers possess all the competencies specified in the Goals for Driver Education (GDE) model(Hatakka et al., 2002), which the Swedish driving curriculum is based upon. The GDE matrix (Table 1), a summary of the conclusions from Goals for Driver Education, lists not only necessary skills and knowledge but also focuses on what risks drivers need to be aware of and on the self-assessment skills needed to be able utilize their skills and knowledge in a safe manner.

Table 1. Goals for Driver Education, adapted from Hatakka et al. (2002).

Level	Knowledge and skills	Risk-increasing factors	Self-evaluation
Goals for life and skills for living (general)	Lifestyle Age Gender/Sex Group Personality	Sensation seeking Peer pressure Youth Disease and disabilities	Drivers own conditions Impulse control Ability to self-reflect
Goals and context of driving (trip related)	Strategic choices of: - Mode of travel - Time of travel - Route - Purpose of travel When? Where? How?	Alcohol Fatigue Competition Darkness Winter road condition Rush hour	Own motive of strategic choices Self-critical thinking
Mastery of traffic situations	Traffic situations Traffic Rules Cooperation Hazard Perception	High speed Short distance Children in traffic Low visibility Wild animals	Calibration of own driving skill Self-critical thinking
Vehicle manoeuvering	Maneuvering Technology Laws of nature	Bad tires No seatbelt Worn breaks	Calibration of own maneuvering skill Motivation to use safety systems

Traditionally, both the driver education and driver test have been focusing on the knowledge and skills of the first two levels of the GDE matrix (Forward et al., 2017; Hatakka et al., 2002, see Table 1). Driving simulator training offers new possibilities to teach novice drivers the competencies contained in higher levels of the GDE matrix, such as strategical hazard perception, risk perception and decision-making, in both the immediate situation up to the whole trip as described in the second and third level of the GDE matrix. Insight based learning has shown promise in also reaching the fourth level (Gregersen, 1996; Senserrick et al., 2001; White et al., 2011) and using simulators as tools to facilitate this learning is promising.

The potential benefits of using driving simulators need to be examined in a rigorous manner and connected to models of driver behavior and theories of cognition. Both the GDE model and theories like Risk Allostasis acknowledge how our goals, motive, attitudes, executive function and metacognition influences the driver's decision-making. Especially young and novice drivers seems to be especially vulnerable to risky decision-making, explained in part by inexperience and maturity (Bates et al., 2019).

The author's PhD project aims to improve driver education with the appropriate use of driver simulators in parts of the education where they can make a difference. The scope of the PhD project is to examine the Swedish driving curriculum with a focus on the GDE model, and to use cognitive science theory and methods to evaluate the necessary skills and knowledges needed to be a safe driver. The driving education and driver assessment has recently been critiqued for not fully teaching and testing the Swedish driving curriculum (Forward et al., 2017). To make sure that the drivers possess the competencies of the GDE-matrix, there is a need to make sure that the drivers get to learn and can practice these competencies, but also that assessment can be made of whether drivers are competent enough to drive in a safe manner.

A study exploring how targeted necessary driving skills from the GDE matrix can be assessed using a driving simulator screening test has been conducted (Thorslund et al., in press). A total of 70 study participants were presented with 16 traffic scenarios in the simulator, which all involved hazards (see Figure 1 for example). These 16 scenarios were created in dialogue with driver examiners from The Swedish Transport Administration and other experts. The tasks in the scenarios require the participant to perceive (potential) hazards and act in a safe way. The screening test differs from a traditional hazard perception test in the level of situational awareness (Horswill, 2016) required and in that test, subjects must act based on the perceived hazard (Jackson et al., 2009).

The results of the study showed that 63% of the participants failed the screening test by performing below the safety threshold set by the experts in at least one of the test scenarios, and that 66% of the failed participants still passed their on-road driving test a few days later. These results can be explained by the different strengths and weaknesses of the two tests. While the on-road driving test attains higher ecological validity and primarily focuses on the knowledges and skills of the two first levels of the GDE matrix, the simulator test has a higher degree of reliability and assesses driver performance in safety-critical situations, mostly on the second level of the GDE matrix. This indicates that a driving simulator screening test can be used as a complement to the knowledge exam and on-road driving test to find drivers who lack necessary competencies to drive safely.

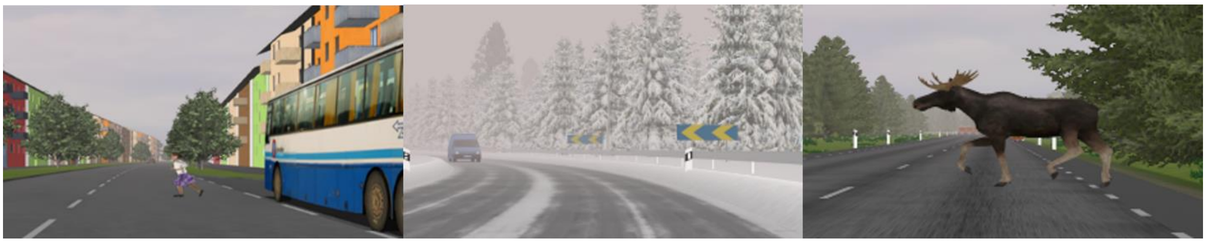


Figure 1. Example of hazards from the screening test.

To move forward the author aims to answer the following research questions during their PhD:

- Can we train young drivers' risk perception with a driving simulator?
- Can we train novice drivers' risk perception with a driving simulator?
- Can we train young drivers' ability to make safe decisions on the road using a driving simulator?
- Can we train novice drivers' ability to make safe decisions on the road using a driving simulator?
- Can we test young novice drivers' risk perception with a driving simulator?
- Can we test young novice drivers' decision-making with a driving simulator?

To achieve this the author plans to conduct empirical studies on education and driver assessment in driving simulators.

References

Bates, L., Rodwell, D., & Matthews, S. (2019). Young driver enforcement within graduated driver licensing systems: a scoping review. *Crime Prevention and Community Safety*, *21*(2), 116–135. https://doi.org/10.1057/s41300-019-00061-x

Directorate-General for Mobility and Transport. (2023, February 21) *Road safety in the EU: fatalities below pre-pandemic levels but progress remains too slow* https://transport.ec.europa.eu/news-events/news/road-safety-eu-fatalities-below-pre-pandemic-levels-progress-remains-too-slow-2023-02-21_en

Forward, S., & Lewin, C. (2006). *Medvetna felhandlingar i trafiken: En litteraturundersökning*. *[Traffic violations: literature review and analysis]* https://www.diva-portal.org/smash/record.jsf?pid=diva2:675273

Forward, S., Nyberg, J., Gustafsson, S., Petter, N., & Henriksson, G. P. (2017). *Den svenska förarutbildningen: dagsläge och framtidsutsikter*. *[Novice driver training in Sweden- present and future prospects]* https://www.diva-portal.org/smash/record.jsf?pid=diva2:1096681

Gregersen, N. P. (1996). Young drivers' overestimation of their own skill - an experiment on the relation between training strategy and skill. In *Accid. Anal. and Prev* (Vol. 28, Issue 2).

Hatakka, M., Keskinen, E., Gregersen, N. P., Glad, A., & Hernetkoski, K. (2002). From control of the vehicle to personal self-control; broadening the perspectives to driver education. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(3), 201–215. https://doi.org/10.1016/S1369-8478(02)00018-9

Horswill, M. S. (2016). Hazard Perception in Driving. *Current Directions in Psychological Science*, *25*(6), 425–430. https://doi.org/10.1177/0963721416663186

Jackson, L., Chapman, P., & Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics*, *52*(2), 154–164. https://doi.org/10.1080/00140130802030714

Keskinen, E. (2007). What is GDE all about and what is it not. *The GDE-Model as a Guide in Driver Training and Testing.* Ingår i Proceedings from the conference. The GDE-model as a guide in driver training and testing. Umeå, may 7–8, 2007. Widar, Henriksson, Tova Stenlund, Anna Sundström, Marie Wiberg- EM No 59, 2007. Umeå: Umeå universitet.

Senserrick, T. M., Swinburne, G. C., & Monash University. Accident Research Centre. (2001). *Evaluation of an insight driver-training program for young drivers*. Monash University Accident Research Centre.

The National Board of Health and Welfare. (n.d.) *Statistikdatabas för dödsorsaker [Statistical database for causes of death]* The National Board of Health and Welfare. Retrieved June 14, 2023, from https://sdb.socialstyrelsen.se/if_dor/val.aspx

Thorslund, B., Thellman, S., Nyberg, V., & Selander, H. (in press). Simulator-based driving test prescreening as a complement to driver testing – toward safer and more risk-aware drivers. Manuscript in press.

White, M. J., Cunningham, L. C., & Titchener, K. (2011). Young drivers' optimism bias for accident risk and driving skill: Accountability and insight experience manipulations. *Accident Analysis and Prevention*, *43*(4), 1309–1315. https://doi.org/10.1016/j.aap.2011.01.013

Reviewer 1. Claes Strannegård, Department of Applied Information Technology, University of Gothenburg

This contribution is a sketch of a plan for a PhD project. The topic of the project is cognitive aspects of safe driving. The author reports results from a partially described screening test, where drivers are shown computer renderings of hazardous situations. Examples of such hazardous situations are pedestrians or moose suddenly appearing in front of the car, or road bends with patches of ice. It was noted that many drivers scored low on this screening test, although they passed the driver's license test only days later. This raises the very relevant question of whether computer simulations of dangerous situations should be included in the education and testing leading up to driver's licenses.

The PhD project seems to be well chosen and well-motivated, but very little substance of relevance to cognitive science is presented in this text. The topic itself is relevant to cognitive science, however, since it combines computer simulations with cognitive psychology and touches upon topics such as attention, perception, memory, and decision-making.

Suggestions:

Please provide more details about the screening test. Further comments can be found in a separate document.

Reviewer 2. Simon Skau, Institute of Neuroscience and Physiology, Gothenburg University

Review of Simulators for driver training and assessment

Overall, these pages lack any discussion of cognition. While you mention your intent to work with cognitive theories and methods, it remains unclear which specific theories and methods you will employ, and how you plan to do so. It's important to elaborate on these points and outline expected outcomes.

In Table 2, the entries only pose questions like "What cognitive skills are here?" without providing any evaluative content. However, I understand that this project is in its early stages, and until the cognitive aspect is thoroughly integrated, an evaluation of the work isn't feasible.

Moreover, after reviewing this, I'm uncertain whether your PhD project will remain purely theoretical—focusing on curriculum and matrices with various cognitive theories—or if empirical work will also be involved. For instance, it's unclear whether you plan to validate that the simulator effectively assesses the cognitive skills you propose it will evaluate. If empirical work is indeed part of your plan, how would you design and execute such research?

Regarding the phrase "focusing on the lower left levels of the GDE matrix," it might be clearer to express it as "Knowledge and Skills for Vehicle" if that's your intended meaning. There's ample space to avoid ambiguity. You mention, "Newer technologies, like driver training simulators, offer new possibilities to teach novice drivers the skills and knowledge contained in higher levels of the GDE matrix." If by "higher" you refer to the levels above the bottom row, it's important to provide clarification on this point. I'm having difficulty understanding how simulators can effectively convey skills and knowledge related to Lifestyle, Age, Gender/Sex, Group, or Personality.

You also state, ”Simulators may also offer new possibilities to ensure that the drivers possess the necessary competencies, especially the higher rightmost competencies of the matrix.” Here, you refer to self-evaluation of the driver’s conditions, impulse control, and self-reflection abilities. Could you provide concrete examples to illustrate this concept?

Towards precision cognitive neuroscience with ecologically valid tasks: a pilot of dense sampling and functional connectivity with fNIRS:

William Hedley Thompson1,2, Victoria Collier1, Amitis Samadi1, Simon Skau3,
1 Department of Applied Information Technology, Gothenburg University, Sweden
2 Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden
3 Institute of Neuroscience and Physiology, Gothenburg University, Sweden
william.hedley.thompson[at]gu.se,

Dense sampling in neuroimaging has gained popularity, particularly in fMRI. In this approach, one or a few participants record brain activity multiple times, sometimes exceeding 100 sessions, to observe within-subject fluctuations (Gordon et al., 2017; Poldrack et al., 2015). These projects have provided valuable insights into the brain's functional organisation (Gordon et al., 2017), psychiatric research (McGowan et al., 2022), brain activity fluctuations during the menstrual cycle (Pritschet et al., 2020), and some have argued their need in addiction research (Yip & Konova, 2022). It is a promising approach to track cognitive fluctuations within individuals relating to, for example, mood, sleep quality, and stress.

All the aforementioned examples involve fMRI. Each non-invasive neuroimaging technique has its own advantages and disadvantages. Functional near-infrared spectroscopy (fNIRS) is an optical imaging technique that measures changes in oxygenated and deoxygenated haemoglobin concentrations. While fNIRS can only measure a couple of centimetres into the cortex, it is less sensitive to movement and can be used in situations where more movement is inevitable, such as with children, patients, or certain tasks.

Here we will perform a pilot for dense sampling with fNIRS. We investigated whether functional connectivity is influenced by: i) the internal state of the individual (caffeine vs. non-caffeine), ii) the task state (problem-solving task vs. rest), and iii) channel placement (correctly placed cap vs. offset cap). The first two variables aim to determine whether fluctuations in the internal state and cognitive task differences are identified—such fluctuations are what a larger dense sampling study would aim to track in a single subject. The final variable aims to evaluate the impact of an offset fNIRS cap, which may result in slightly different coverage—a possibility during larger data collection that channels might have some experimental noise due to placement. In addition to dense sampling, we aim to assess functional connectivity and network analyses of fNIRS data. While these analysis strategies are commonly applied in fMRI, they are relatively less frequently utilized in fNIRS studies (e.g. Skau et al., 2022). This pilot study will allow us to assess the appropriateness and interpretability of employing these analysis strategies in the context of fNIRS, dense sampling and cognitive tasks.

Method

Data. The experiment took place over 10 days at the Department of Psychology in Gothenburg. The participant was a 35-year-old healthy male. He was informed whether they could drink coffee 45 minutes before each testing session, which was randomised. The participant wore an EasyCap with fNIRS optodes. Each day, five recordings were conducted, all in a dark environment. The first recording was a resting state (correctly placed), a second resting state (the cap shifted backwards by 20 mm) and a third (correctly placed). Each resting state was 5 minutes. The fourth and fifth recordings involved the participant playing Sudoku for 5-minute blocks in each (fourth: cap correctly placed, fifth: cap shifted background 20mm). The sudoku level was "hard", and the puzzle was not completed within the time limit. The sudoku was performed on the participant's mobile phone using an app. In a total of 50 separate recordings (5 recordings for each session). Note: the participant is a co-author of this paper and was involved in the design. The study was approved by the Swedish Ethical Review Board (Dnr: 2022-01702-01).

Preprocessing. Preprocessing of fNIRS data was done with the MATLAB package HomER2 (Huppert et al., 2009). The raw data were converted to optical density, and a band-pass filter was applied (0.05 - 0.5Hz) to remove drift and respiration artefacts. Motion artefacts were corrected with a Cubic spline filter. Conversion from optical density to haemoglobin concentration was done using pathlength

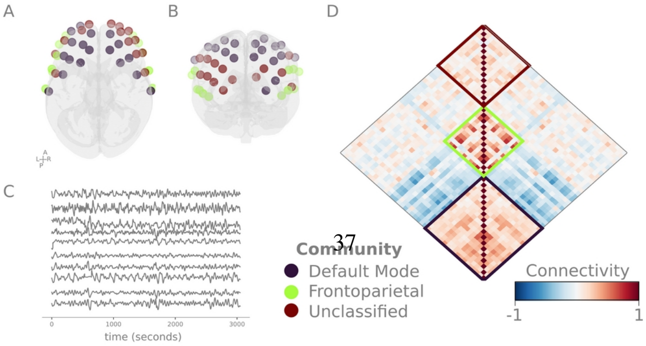


Figure 1: Shows the coverage of the 44 channels. A above the brain and B from the front. C. Example of recorded time series D. Connectivity matrix of 44x44 correlations averaged over all 50 sessions.

factors of 6.4 and 5.4 for both wavelengths. Activity from short separation channels was regressed out of the 44 standard channels. Only oxy-Hb was used for analysis. Finally, each channel was demeaned by subtracting the average global signal.

Data analysis. For each run, a 44x44 network was constructed by pairwise Pearson correlations (see Figure 1). We analysed global-level metrics (i.e. one measure per network) and edge-level metrics (differences in clusters of edges). We chose modularity and global efficiency for the global measures of the network's organization. Global efficiency is the inverse average shortest path which entails that it is theoretically possible to travel across the network in fewer steps. Modularity estimates how the strength of between-community edges after deriving the community parcellation that maximises modularity (Rubinov & Sporns, 2010). Cluster statistics were run on the edge level analyses, where the 44x44 networks obtained per run are compared between conditions. All runs were analysed together.

The Louvain algorithm (Blondel et al., 2008) identified three communities. To identify these communities, we found the closest Yeo 7 network based on the nearest parcel from Schaefer 2018 (Schaefer et al., 2018; Yeo et al., 2011) to the estimated MNI coordinates of each channel. Based on the majority of nodes in each community, we identified the default mode network, frontoparietal network, and a mix of nodes from the frontoparietal, motor and attention networks (see Figure 1). Visualizations were made in *NetPlotBrain* v0.3.0 (Fanton & Thompson, 2023).

Statistics. To analyze the three different levels, the following statistical tests were conducted. Two separate 2x2 within-subject ANOVA with *Caffeine* (yes, no) and *Task* (sudoku, rest) as the independent variables and the global measures as the dependent variables. For edge legal analyses, network-based statistics (Zalesky et al., 2010), a non-parametric method for identifying clusters of edges was run separately for the caffeine and task conditions. The cluster statistics were set as 3.5, with 10000 permutations and alpha = 0.001, two-tailed. Importantly, clusters should not be interpreted as the collection of significant edges, but that there is some cluster is different from the null hypothesis.

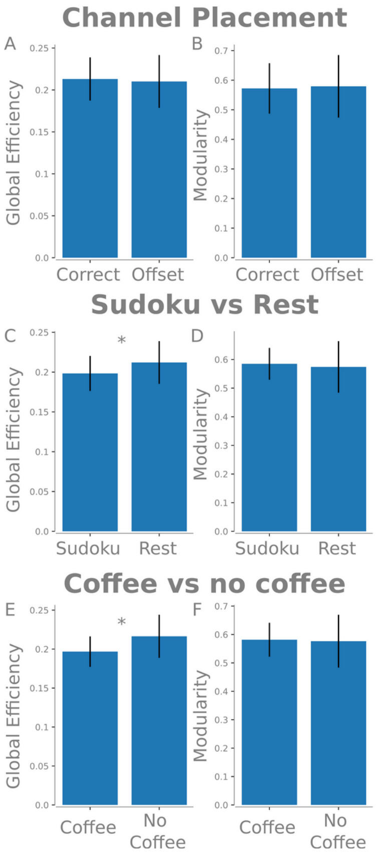


Figure 2: Global efficiency and modularity for three different experimental conditions. Channel placement for A Global efficiency and B Modularity; Coffee: C Global efficiency and D Modularity. Sudoku vs rest E Global efficiency F Modularity.

Results

Global-level results. We found no difference due to the placement of channels (correct, offset) for either measure (*Global efficiency*: T=0.257, p=0.801; *Modularity*: T=-0.165, p=0.871). The same applied to the sudoku task correct and incorrect placements (*Global efficiency*: T=0.949, p=0.355; *Modularity*: T=-0.333, p=0.743) Given this, we pooled sessions regardless of offset for all analyses (see *limitations*). We found a difference in the global efficiency for both *Task* (sudoku, rest) and *Coffee* (yes, no) with F=5.175, p=0.0467, η2=0.083 and F=8.875, p=0.0046, η2=0.162 respectively, with no significant interaction between the two (F=2.487, p=0.122, η2=0.513). Modularity showed no significant effect for *Task* (F=0.251, p=0.618, η2<0.005), *Coffee* (F=0.030, p=0.864, η2<0.001) or interaction (F=0.002, p=0.966, η2<0.001).

Edge-level results. One cluster was found for *Task* (p<0.001, Figure 3BCD). Here we see that within-community edges are higher for rest, and many between-community edges also decrease. This explains why the global efficiency decreases for sudoku. However, we also see that for sudoku, there is an increase in between community edges between the default mode network and the frontoparietal network (Figure 3A). This increase in between-network connectivity for sudoku and decrease in within-community edges was part of the significant cluster (Figure 3BCD). For *Coffee*, a similar result was found with a single cluster (p<0.001, Figure 3D-G).

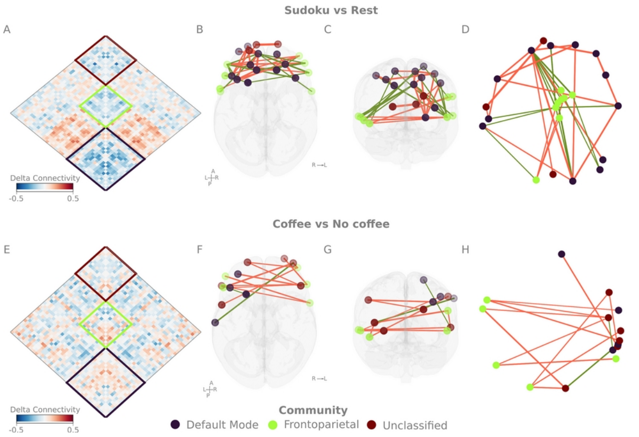


Figure 3: Edge-level differences in sudoku vs rest (Top) and coffee vs no-coffee (bottom). A: Difference in the connectivity matrix between conditions, red indicates greater for sudoku, BC: the significant cluster; green edges indicate sudoku > rest, red shows edge where rest > sudoku. D: Spring layout visualization of the cluster. E-G save as A-D, but for coffee vs no coffee

Discussion and Conclusion

The results of this exploratory analysis of a dense sampling pilot in fNIRS indicate that it is a promising approach for detecting cognitive and state fluctuations over time in a larger experiment. We observed significant differences in global networks and patterns of edge clusters related to both manipulations. Interestingly, in the case of the difference between the sudoku task and rest, decreased global efficiency was accompanied by a general decrease in connectivity during the sudoku task, except, importantly, for the increased integration between the default mode network and frontoparietal network. Precious studies using fNIRS and sudoku found increased prefrontal activation based on sudoku difficulty (Ashlesh et al., 2020). Further, functional connectivity research has found that integration and segregation of brain networks relates to different cognitive tasks (e.g. Cohen & D'Esposito 2016). Such cooperation between the frontoparietal and default mode networks is expected here since their cooperation has been suggested to be needed for internal thought (Smallwood et al., 2012) – which is a likely process during the sudoku task. A similar global difference was observed for the coffee manipulation, which was attributed to a decrease in between-community connections. Overall, this study demonstrates that functional connectivity analysis strategies can detect differences in dense sampled fNIRS data for both cognitive and state manipulations. However, it is important to note that these results are based on a pilot study, and therefore, they should be interpreted as exploratory, especially since certain decisions (e.g., pooling) were made after data collection.

Limitations. Pooling data from trials where the cap has been offset may not be advisable in future recordings. Nevertheless, this study showed that small perturbations of a few millimetres should not have a significant impact on network analyses. Another important consideration is that cap shifting is likely to have less impact on global measures compared to edge-based measures. Global properties should remain relatively consistent regardless of cap placement, whereas specific edges will have variance in their relationships.

References

Ashlesh, P., et al. (2020). Role of prefrontal cortex during Sudoku task: fNIRS study. *Translational neuroscience*, *11*(1), 419-427.

Blondel, V. D., et al, (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10008(10), 6.

Cohen, J. R., & D'Esposito, M. (2016). The segregation and integration of distinct brain networks and their relationship to cognition. *Journal of Neuroscience*, *36*(48), 12083-12094.

Fanton, S., & Thompson, W. H. (2023). NetPlotBrain: A Python package for visualizing networks and brains. Network Neuroscience, 7(2), 461–477.

Gordon, E. M., et al (2017). Precision Functional Mapping of Individual Human NeuroResource Precision Functional Mapping of Individual Human Brains. Neuron, 1–17.

Huppert, T. J., et. al. (2009). HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. Applied Optics, 48(10)

McGowan, A. L., et al (2022). Dense Sampling Approaches for Psychiatry Research: Combining Scanners and Smartphones. Biological Psychiatry.

Poldrack, R. A., et al (2015). Long-term neural and physiological phenotyping of a single human. Nature Communications, 6.

Pritschet, L., et al (2020). Functional reorganization of brain networks across the human menstrual cycle. NeuroImage, 220, 117091.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. Neuroimage, 52(3), 1059-1069.

Schaefer, A., et al (2018). Local-Global Parcellation of the Human Cerebral Cortex From Intrinsic Functional Connectivity MRI. Cerebral Cortex, 28, 3095–3114.

smallwood, J., et. al. (2012). Cooperation between the default mode network and the frontal–parietal network in the production of an internal train of thought. *Brain research*, *1428*, 60-70.

Skau, S., et. al. (2022). Segregation over time in functional networks in prefrontal cortex for individuals suffering from pathological fatigue after traumatic brain injury. Frontiers in Neuroscience, 16, 972720.

Yeo, B., et. al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of Neurophysiology, 106, 1125–1165.

Yip, S. W., & Konova, A. B. (2022). Densely sampled neuroimaging for maximizing clinical insight in psychiatric and addiction disorders. Neuropsychopharmacology, 47(1),

Zalesky, A., et al.,. T. (2010). Network-based statistic: Identifying differences in brain networks. NeuroImage, 53(4), 1197–1207.

Reviewer 1. Sara Stillesjö, Department of Psychology, University of Umeå

The authors have touched upon a timely and very interesting topic that expands the horizon on how fNIRS can be used to better understand the brain´s functional organization. To summarize, the authors have conducted a pilot study that investigates the use of fNIRS for dense sampling of data from the same participant during performance of a cognitive task, rest, and the internal state of the individual. They demonstrate that this approach to evaluate functional connectivity analysis strategies can detect differences in separate brain networks from different cognitive and state manipulations.

Some sections in the manuscript could benefit from further clarification, and I hope that my comments can help to increase the quality of the paper.

The experimental design used for the sudoku recordings is somewhat vague. Given that the authors contrast brain networks involved in task and rest, it is important to use a methodological approach that allow for a separation of blood flow associated with performance of the cognitive task from other processes. A block design or an event-related design that alters task and rest (to let the signal return to baseline) is preferable. If the design doesn´t control for this, the sudoku data will likely include plenty of noise from other internal and external sources, and this would compromise the results. Thus, readers would benefit from a more detailed description of the experimental design to be able to evaluate the quality of the results.

It is not clear why channel placement is included as a question of interest, and where the assumption that channel placement can affect functional connectivity stems from. This can be better motivated with scientific references and/or an elaborated discussion.

The authors write that the channel placement was analyzed as the global average across task and rest conditions (correct vs incorrect placement). For transparency, it would be informative to see the comparisons for task and rest displayed in a table/graph as well.

Results related to the one significant cluster and community edges should more precisely state where in the brain/ what network the results were observed.

The discussion would benefit from a more extensive elaboration on what the results imply and anchored in up-to-date references related to functional brain networks, how they support different cognitive processes, and cooperate during different (cognitive) states to put the results in context.

Reviewer 2. Linnea Karlsson – Wirebring, Department of Psychology, University of Umeå

This paper by Hedley Thompson and colleagues describes a pilot study with the aim of exploring the potential of using dense sampling and functional connectivity to detect cognitive and state fluctuations in fNIRS data (functional near-infrared spectroscopy).

The study is very ambitious, including a total of 50 separate fNIRS recordings (five 5-minute sessions per day over 10 consecutive days). The study is also novel in scope, as the methodological approach presented has been more thoroughly tested on fMRI- compared to fNIRS-data.

Here are some comments:

Dense sampling is resource-consuming. Can the authors add a few lines to the introduction exemplifying the general benefits of dense sampling of neuroimaging data such as fNIRS? Can the authors clarify how dense sampling helps in achieving the aims of the current study, in the section where the aims are described? It would be informative with a theoretical argument for why caffeine vs non-caffeine was chosen as manipulation of "internal state". Can it be expected a priori to interact with the Task-variable (Sudoku) on the measures chosen? The authors suggest and use three different measures: global efficiency, modularity and edge-level analyses. Why were these measures chosen? Can these measures be expected to be equally reliable a priori? Are there any potential biases in these measures? The "offset cap" condition come across as rather odd, as it is likely to introduce some measurement noise. The authors acknowledge that pooling the data across the offset conditions might be a limitation of the study. Was the data pooled also for the edge-level analysis? How was that motivated? What motivated the choice to have three Rest-sessions and two Task-sessions instead of three of each? In the discussion section, the authors interpret their results as "increased integration between the default mode network and frontoparietal network". This is arguably a somewhat bold claim. Are there any alternative interpretations? The authors assert that the results of the current pilot study indicate that dense sampling of fNIRS data is a promising approach. However, in the current version of the ms, the authors do not present an evaluation of the benefits of dense sampling in this context. Comparisons with other dense-sampled individuals would be one way to go about. It would also be informative to see some analyses of the measures of interest on the time dimension, which I believe would be possible using the already collected pilot data. For example, would the same differences in global efficiency between Sudoku and Rest be observed already after the third day?

Reply to reviews

Reviewer 1.

The authors have touched upon a timely and very interesting topic that expands the horizon on how fNIRS can be used to better understand the brain´s functional organization. To summarize, the authors have conducted a pilot study that investigates the use of fNIRS for dense sampling of data from the same participant during performance of a cognitive task, rest, and the internal state of the individual. They demonstrate that this approach to evaluate functional connectivity analysis strategies can detect differences in separate brain networks from different cognitive and state manipulations.
Some sections in the manuscript could benefit from further clarification, and I hope that my comments can help to increase the quality of the paper.

General Reply: we are grateful for the comments. Unfortunately, since the short paper format wants only a 3-page article, some of the wider discussion points could not be added.

Comment R1.1
The experimental design used for the sudoku recordings is somewhat vague. Given that the authors contrast brain networks involved in task and rest, it is important to use a methodological approach that allow for a separation of blood flow associated with performance of the cognitive task from other processes. A block design or an event-related design that alters task and rest (to let the signal return to baseline) is preferable. If the design doesn´t control for this, the sudoku data will likely include plenty of noise from other internal and external sources, and this would compromise the results. Thus, readers would benefit from a more detailed description of the experimental design to be able to evaluate the quality of the results.

Reply: We understand this was a little unclear. The design here is a block design. We have added the word blocks to avoid confusion. During the 5-minute Sudoku block, the participant only focused on solving that. This is to try and give a more naturalistic problem-solving task compared to rest. In addition, short separation channels were used, thus enabling us to regress out any added noise from the scalp and skull that could have been introduced by the task.

Comment R1.2
It is not clear why channel placement is included as a question of interest, and where the assumption that channel placement can affect functional connectivity stems from. This can be better motivated with scientific references and/or an elaborated discussion.

Reply: We understand this was a little unclear. This was one of the motivations for doing the pilot. Since fNIRS allows for more movement, it is possible that the cap shifts during recording or that the cap may be slightly offset during multiple recordings (not to as large a degree as here, but there may be smaller ~1mm shifts every day in different directions). When recording multiple people, this is not a problem. However, if the data is coming from the same person, this might add up to some considerable noise. So one of the reasons for doing this pilot was to check, especially when looking at large-scale network properties, how this impacts the results. Obviously, if you want to record a localized spot, then such an offset will have a greater impact.

Comment R1.3
The authors write that the channel placement was analyzed as the global average across task and rest conditions (correct vs incorrect placement). For transparency, it would be informative to see the comparisons for task and rest displayed in a table/graph as well.

Reply: The reviewer is correct that there is no good reason why we did not include that information - our reasoning was that it was not in the figure. That has been added to the text.

Comment R1.4
Results related to the one significant cluster and community edges should more precisely state where in the brain/ what network the results were observed.

Reply: We understand why the reviewer wants this. However, one pitfall of cluster-based statistics is to misinterpret the “significant cluster”. All we can say is that there is a significant cluster, but we cannot say what the exact size is. This is why we have presented the delta connectivity matrix and interpreted the results broadly relating to how the different brain networks appear to be interacting (both within the “significant cluster” but also consistent outside) instead of relating to exactly which brain areas. Our motivation for this? See for example, the excellent explanation on the field trip toolbox for how to not interpret cluster-based permutation tests (although here it is for temporal-spartial-frequency clusters for EEG/MEG the principle holds here): https://www.fieldtriptoolbox.org/faq/how_not_to_interpret_results_from_a_cluster-based_permutation_test/

Comment R1.5
The discussion would benefit from a more extensive elaboration on what the results imply and anchored in up-to-date references related to functional brain networks, how they support different cognitive processes, and cooperate during different (cognitive) states to put the results in context.

Reply: We agree that this could have been a little longer to discuss and that the results are in line with litear Since the paper was only 3 pages long. We have added a slightly longer discussion (which hopefully also gives a little more detail for R1.4) - (which means we have unfortunately gone over the page limit).

Reviewer 2.

This paper by Hedley Thompson and colleagues describes a pilot study with the aim of exploring the potential of using dense sampling and functional connectivity to detect cognitive and state fluctuations in fNIRS data (functional near-infrared spectroscopy).

The study is very ambitious, including a total of 50 separate fNIRS recordings (five 5-minute sessions per day over 10 consecutive days). The study is also novel in scope, as the methodological approach presented has been more thoroughly tested on fMRI- compared to fNIRS-data.

General Reply: Thank you for the review. All useful comments. We have done our best to explain our motivation behind several of our decisions in a little more detail.

Comment R2.1
Dense sampling is resource-consuming. Can the authors add a few lines to the introduction exemplifying the general benefits of dense sampling of neuroimaging data such as fNIRS?

Reply: We are not entirely sure if we agree. Dense sampling is *demanding* on the several participants that are a part of the study, but if you record 2 people 50 times or 100 people 1 time, there is not much difference in resources consumed. The only difference is that those two people must be part of the experiment for a longer time (so finding those people who are willing is a potential challenge). However, we have added a sentence about what dense sampling can offer.

Comment R2.2
Can the authors clarify how dense sampling helps in achieving the aims of the current study, in the section where the aims are described?

Reply: We have amended a sentence when we discuss our aims, which now reads:

> “The first two variables aim to determine whether fluctuations in the internal state and cognitive task differences are identified—such fluctuations are what a larger dense sampling study would aim to track in a single subject.”

Comment R2.3

It would be informative with a theoretical argument for why caffeine vs non-caffeine was chosen as manipulation of "internal state". Can it be expected a priori to interact with the Task-variable (Sudoku) on the measures chosen?

Reply: In the larger study, we would track natural fluctuations (mood, sleep quality, stress, etc.). However, for the limited number of recordings for the pilot, we wanted to directly manipulate the "internal state" because we would not know if there would be enough natural fluctuations of sleep or mood. Manipulating whether the participant had drunk coffee or not was a simple way to ensure there was enough data where there was some manipulation of the subjective cognitive state of the participant. It is possible that coffee interacts with problem-solving/mathematical reasoning. That is why we checked (and reported) if there was an interaction, and there was not. In addition, *caffeine* has been shown to affect the hemodynamic response during cognitive tasks with fNIRS, so it was a logical choice there too. However, none of these studies investigate the effect of *caffeine on functional connectivity.*

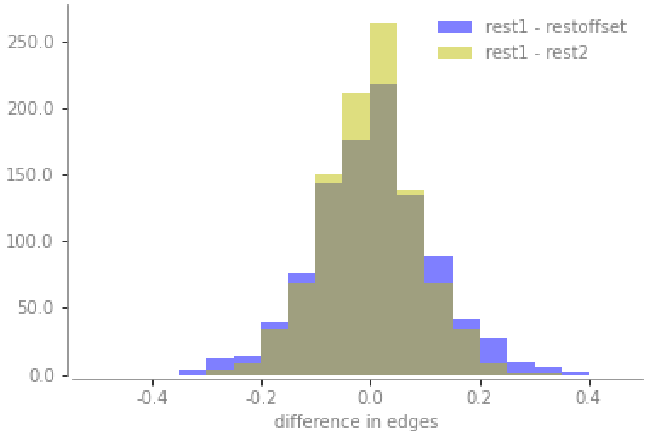
Comment R2.3
The authors suggest and use three different measures: global efficiency, modularity and edge-level analyses. Why were these measures chosen? Can these measures be expected to be equally reliable a priori? Are there any potential biases in these measures?

Reply: We chose two global network measures that quantify different topographical properties: one is based on shortest paths (efficiency), and the second measures the denseness of connections (modularity). Both are used frequently in neuroimaging and in network theory. Others could have been chosen that quantify similar underlying properties of the network.

Comment R2.4
The "offset cap" condition come across as rather odd, as it is likely to introduce some measurement noise. The authors acknowledge that pooling the data across the offset conditions might be a limitation of the study. Was the data pooled also for the edge-level analysis? How was that motivated?

Reply: For the pilot, we wanted to see how much this induced noise impacts the results. We considered this an important methodological issue as there may be (less) noise in the full study where the cap might have minor (~1mm) differences in placement a day. Thus it is important to test what impact such noise has on the data in this pilot. This was one of the issues we wanted to check within the pilot. (See Reply 1.2 for more motivation).

A motivation for pooling the edge data was that the connectivity matrices were relatively similar between the two conditions. While we admit we do not think this is an ideal step, we included this here. We checked after this review and,if you only look at, for example, the resting state and sudoku with the "correct" placement, then there is still a significant cluster, but, as expected with the less data, the results are less pronounced. The motivation was to include more data despite the known noise relating to the cap manipulation. Another way to justify that it was acceptable to merge the conditions is the following figure (right). Here we see the distribution of the difference in the average connectivity matrices for rest (run 1, correct) with rest (run 2, correct) and rest (offset). Where the difference with the offset condition has some more extreme differences in connectivity, it shows that we have not pooled over sessions that are considerably different than each other.



Comment R2.5
What motivated the choice to have three Rest-sessions and two Task-sessions instead of three of each?

Reply: When designing the pilot, we tried to outline a number of unknowns that we hoped the pilot could answer before a larger data collection. The original plan was three Rest and one Task. The second task with an offset condition was added at the last minute so we could see the impact of cap placement there (giving us even more data to evaluate how much noise is induced from cap placement). If we were redesigning the pilot today, this would be two or three of each, absolutely. We do not plan to have multiple tasks in the full experiment.

Comment R2.6
In the discussion section, the authors interpret their results as "increased integration between the default mode network and frontoparietal network". This is arguably a somewhat bold claim. Are there any alternative interpretations?

Reply: We are not entirely sure what is bold about this claim, especially considering Figure 3A shows an increase in the connections between the frontoparietal network and default mode network - and parts of these nodes and connections are found in the significant cluster (Figure 3B-D). However, we have added more text to the discussion, which should explain this a little more (see reply 1.5 discussion)

Comment R2.7
The authors assert that the results of the current pilot study indicate that dense sampling of fNIRS data is a promising approach. However, in the current version of the ms, the authors do not present an evaluation of the benefits of dense sampling in this context. Comparisons with other dense-sampled individuals would be one way to go about. It would also be informative to see some analyses of the measures of interest on the time dimension, which I believe would be possible using the already collected pilot data. For example, would the same differences in global efficiency between Sudoku and Rest be observed already after the third day?

Reply: We do not think we make this strong a claim. We qualitfy the statement: "fNIRS indicate that it is a promising approach *for detecting cognitive and state fluctuations in a larger experiment*". We have added the words "over time" to this statement in order to make it clearer. Further, in the discussion and conclusion, we state: " Overall, this study demonstrates that functional connectivity analysis strategies can detect differences in dense sampled fNIRS data for both cognitive and state manipulations." We think that is what this pilot shows. The additional analyses that the reviewer suggested that we are already considering in the full study are:

1. Multiple dense sampled subjects.
2. Adding multiple analyses over the time dimension

And we think they are excellent suggestions. (2) will definitely happen in the full study (at present, we do not think there is enough data to do such analysis justice), and we are already considering some redesign to include a few subjects rather than just a single subject.

Category theory: From Semantic Information in the mind to Shannon Information in the brain

Mohammad-Hossein Heidari Beni[1], Mariam Marlen Mirstrom[2], Yousef Javaherian[3], and Mohammad Matin Mazaheri Kohani[4]
[1] Department of Telecommunications and Information Processing, Ghent University
[2] Department of Philosophy, Lund University
[3] Department of Computer Engineering, Isfahan University of Technology
[4] Department of Mathematics, Tehran University
MohammadHossein.Heydari@UGent.be

One of the most challenging questions across philosophy, psychology, and neuroscience is the distinction between the mind and the brain. Bridging this divide has significant implications for psychiatric treatments (Kapur, 2003), understanding mental disorders (Clark & Sahakian, 2022), developing artificial general intelligence (Hassabis et al., 2017), formulating advanced mathematical theories of consciousness (Tononi et al., 2016), and many other areas.

Several significant theories, such as physicalism (Clark, 2000), functionalism (Block, 1996), intentionalism (Dennett, 1983), and reductionism (Churchland, 1982), have been posited to elucidate the relationship between the mind and the brain. Physicalism asserts that the mind does not exist independently but emerges from specific physical properties of the brain. In contrast, reductionism, without emphasizing the dependent nature of the mind, aims to express mental states by reducing them to fundamental brain states. Hence, while both theories attempt to equate mental states with brain states, reductionism poses a methodological perspective rather than an ontological one. Counter to physicalism, functionalism posits that mental states can be exhaustively described by their functional roles and relationships. Similarly, intentionalism acknowledges the mind's independent existence, contending that mental states possess unique intentional contents. Intentionalism underscores the intent of these states by emphasizing the representative nature of mental states. Both intentionalism and functionalism's approaches align with information-based solutions to the mind-body quandary. The information-based stance is a subset of functionalism, wherein functions concern information processing, and mental states maintain informational connections. Moreover, intentionalism defines the mental states by the intention to represent things or give information about things. Therefore, it postulates a representational framework for the mind consistent with informational architecture underpinning mental representation.

At the mind level, information processing shapes meanings or concepts as mental representations (Ramos, 2014). At the brain level, information processing is derived from statistical patterns of neural activities and focuses on shaping these statistical patterns spatially or temporally. Due to the vital role of information in characterizing phenomena at both levels, it is essential to find a meaningful connection between the two to bridge the gap between the mind and brain. One possible perspective is that the gap stems from applying two distinct mathematical fields to define information at each level. For instance, the information theory that Shannon developed (1948) proffers mathematical tools to explain neural activities in the brain through their statistical interactions in temporal and spatial domains. Mathematically, neural activities in the brain might also be viewed as random fields (Vanmarcke, 2010) across time and space. Such fields determine the magnitude and structure of information flow over time or between brain regions, which is crucial for cognitive tasks (Sharma et al., 2021).

Information, nevertheless, at the mind level, concerns the representation of concepts or meanings captured in representational structures such as a concept space (Gardenfors, 2004), physical symbol systems (Newell & Simon, 1976), and languages (Proudfoot, 2009). Thus, how exactly the random field of neural activities is related to the representational structures for concepts and meanings at the mind level must be clarified. Regarding the identification of the structural relationship between these two mathematical domains, category theory (Mac Lane, 2013) is a conceivable explanation for studying and relating such structures.

Before diving into how category theory might bridge the gap between Shannon's information theory and representational structures, let us first build an intuition behind their potential connection. At its core, Shannon's information theory quantifies the transmission of information between two random variables. This quantification hinges on the average length of a particular kind of code, often referred to as a sequence of information bits, especially when dealing with binary codes. These sequences encapsulate and convey information within communication frameworks (Shannon, 1948). Nevertheless, Shannon's information theory intentionally ignores these codes' meaning or significance, merely focusing on quantitative analysis of information through a communication channel. When meanings are assigned to these sequences of bits, a specific purpose or function is inherently associated with them. Thus, interpreting such information bits is tied to their utility or role in a task, introducing a challenge: How do we link functionality with raw information? Stated differently, the nature of a task dictates which information is deemed relevant or meaningful and which not. Furthermore, Kolchinsky and Wolpert (2018) state that semantic information is the essential data a system requires about its environment to keep its existence as a task. Adding another layer to this discourse, Wittgenstein introduced the idea of a 'language game.' He claimed that words in a language derive meaning based on their utility or function within a specific game or context. Such games are determined by the challenges or objectives a player confronts to succeed or score (Proudfoot, 2009).

Shannon's definition of information as the message transmitted from a channel's output to its input through a communication system can be expanded to include any connection between input and output, not just limited to communication systems (Shannon, 1948). Furthermore, a task can be seen as a process where the subject works on input to obtain the desired output for optimal utility. The meaning of information arises from defining this relationship between input and output within the context of a task. The collection of tasks or problem space is structurally related to the collection of information, which can be thought of as the information space. The information space encompasses all possible data, facts, or knowledge about a particular subject or context, which can be viewed as a more decadent space that connects the problem space to the solution space. From another perspective, as an examination of the possession of information, meanings, or concepts in an individual's mind, a task or problem might be designed to test their ability to solve it optimally. Solving the problem or optimally performing the task can result in the individual having the information, meaning, or concept in their mind. Therefore, the collection of tasks (task space) is structurally related to information collection. The structure of the task space can relate to the compositional structure of information at both the mind and brain levels. Combining two tasks can result in another task in the same way as combining two concepts or meanings results in another concept. The earlier mentioned compositional structure might be extended to information processed and combined in the brain model described in terms of Shannon information.

Given the conceptualization of information as a relational entity spanning various levels of the mind and neural networks in the brain, category theory might be underscored to bridge information at these levels. At its core, category theory utilizes arrows to model relations in a general view. Conceptually, a category resembles a directed graph enriched with fundamental structures. The directed edges and nodes are designated as arrows and objects within this framework, respectively. A distinctive feature of category theory, setting it apart from a mere directed graph, is the compositional capacity of arrows. Specifically, when the domain of one arrow coincides with the codomain of another, they can be combined to form a novel arrow whose domain and codomain derive from the domain of the first arrow and the codomain of the second arrow, respectively. Each object has an associated identity arrow, marked by congruent domain and codomain. This emphasis on arrows facilitates representing all components, including objects, in arrow-centric terms. The inherent compositional nature of category theory allows it to delineate various structures based on sets representable via arrows. Moreover, this theory introduces the concept of a functor, a mechanism preserving structural relationships between two categories. Preservation implies a mapping that aligns arrows and objects from one category to another while maintaining compositional integrity. In more exact terms, preserving compositional integrity through mapping between two categories means if the composition of arrow A and arrow B results in arrow C in the origin category, the composition of arrow A and B pictures results in the picture of arrow C in the destination category. As such, the foundational principles of compositionality and functor present in category theory offer promising avenues for modeling and interlinking the structures of both task and concept spaces.

As set forth earlier, a task, by assigning significance to a specific relation through creating a distinct functionality for it, brings meaning to that relation. In other words, a task designates a specific input-output relationship as meaningful while treating other relationships as meaningless or irrelevant. Therefore, meaningful information can be characterized as the specific relation necessary for a task. Mathematically speaking, executing a task can be an optimization process over the available relations between inputs and outputs in the context of the task, which determines a specific one as the optimal relation. In category theory, there is a fundamental concept of "universal construction": a unique arrow or a set of arrows with special attributes. Therefore, given optimization as finding a universal construction (Phillips & Wilson, 2016), a task can be expressed as finding a universal construction, a specific arrow with a particular property in category theory, and uniquely determined. To be more precise, the objective function in optimization can be viewed as a functor from one category to the category of real numbers. In this category of real numbers, each object is an actual number. The existence of an arrow from one object to another indicates that the first object is less than the second, representing a specific relation between the two numbers. The fundamental structure of real numbers, based on order, can be depicted through a set of arrows between real numbers, signifying their relative values. In general category theory, ordering a set in this manner is called a 'poset,' as described by Awodey (2006). When a functor from one category to the category of real numbers preserves this structure, it imposes the order on the objects of the initial category in the same manner as in the category of real numbers. Consequently, a terminal object, considered a universal construction type, is a unique object in a category. For this object, there exists a unique arrow originating from every other object in that category. In category theory, this terminal object can be regarded as the optimal solution, marking the result of the optimization process.

In category theory, neurons have been presented in a categorical model, characterized as a set of arrows connecting random variables across time and distinct brain areas, as suggested by Gómez-Ramirez (2014). This model can be considered a categorical version of a random field, Vanmarcke (2010). Within this framework, the objective function is defined as a functor originating from this category or a higher-level category derived from it to the category of real numbers. By defining an order, this function determines a unique arrangement of arrows in the categorical model of neurons, recognized as a terminal object in the high-level category theory, for the associated task. Engaging in a task that inherently modulates a specific random neural field over periods and distinct cerebral zones is analogous to identifying a universal construction as a set of arrows modeling causal interconnections among neural activities, anchored within the context of the objective function - a functor - intrinsically linked with the task. Therefore, if the compositional structure of both concept and task spaces are comprehensively mapped via a functor, it becomes feasible to associate a concept or meaning to a particular neural random field uniquely determined by the corresponding task. In other words, a particular concept can be aligned with unique causal connections or neural random fields, being modeled as a universal construction inherently bound to the relevant task.

Given the insights of Susan Carey (2011), concepts are frequently acknowledged as cognitive representations. These can be detailed as arrows bridging the elements they exemplify and their direct representations. Such a view of concepts or meanings as relations not only exhibits the capability of having the compositional structure in terms of category theory but also has the potential to be synchronized with a task via a universal construction. This perspective elucidates that the mental procedures of crafting and refining meanings find their counterpart in neural activities through identifying the universal construction to characterize a task. The complexity arises when one attempts to represent the task as a functor that extends across a category bound with neural activities and channels it towards a poset category, ensuring the task’s categorical essence harmonizes with another category designated as the conceptual space.

Associating each concept with a stochastic process, as a specific kind of random field, is consistent with the approach of dynamic logic theory (Vityaev et al., 2013). Dynamic logic attempts to connect structural knowledge as a property of logic with dynamics that can effectively model the learning process. Additionally, from the perspective of Integrated Information Theory (IIT) (Oizumi et al., 2014), each concept can result from the causal role of a set of elements, formalized through the random field over elements, which maximizes irreducibility. Therefore, in the perspective of IIT, a concept is also related to a mechanism that results from the optimization of an objective function, which is related to irreducibility or the level of integration.

References

Awodey, S. (2006). *Category theory*. Oxford University Press.

Block, N. (1996). *Functionalism.* In The Encyclopedia of Philosophy supplement. MacMillan.

Churchland, P. S. (1982). *Mind-brain reduction: New light from the philosophy of science*. Neuroscience, 7(5), 1041–1047.

Clark, A. (2000). *Mindware: An introduction to the philosophy of cognitive science*. Oxford University Press.

Clark, L., & Sahakian, B. J. (2022). *Cognitive neuroscience and brain imaging in bipolar disorder*. Dialogues in Clinical Neuroscience.

Dennett, D. C. (1983). *Intentional systems in cognitive ethology: The "Panglossian paradigm*". Behavioral and Brain Sciences, 6(3), 343–355.

Gardenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT Press.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). *Neuroscience-inspired artificial intelligence*. Neuron, 95(2), 245–258.

Howard-Jones, P. (2010). *Introducing neuroeducational research: Neuroscience, education, and the brain from contexts to practice*. Taylor & Francis.

Kapur, S. (2003). *Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia*. American Journal of Psychiatry, 160(1), 13–23.

Kolchinsky, A., & Wolpert, D. H. (2018). *Semantic information, autonomous agency, and non-equilibrium statistical physics*. Interface Focus, 8(6), 20180041.

Mac Lane, S. (2013). *Categories for the working mathematician* (Vol. 5). Springer Science & Business Media.

Newell, A., & Simon, H. A. (1976). *Computer science as empirical inquiry: Symbols and search*. Communications of the ACM, 19(3), 113–126. https://doi.org/10.1145/360018.360022

Oizumi, M., Albantakis, L., & Tononi, G. (2014). *From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0*. PLoS Computational Biology, 10(5), e1003588.

Phillips, S., & Wilson, W. H. (2016). *Systematicity and a Categorical Theory of Cognitive Architecture: Universal Construction in Context*. Frontiers in Psychology, 7.

Proudfoot, D. (2009). *Meaning and mind: Wittgenstein's relevance for the 'Does language shape thought?' debate*. New Ideas in Psychology, 27(2), 163–183.

Ramos, T. (2014). *The concepts of representation and information in explanatory theories of human behavior.* Frontiers in Psychology, 5(1034).

Shannon, C. E. (1948). *A mathematical theory of communication*. The Bell System Technical Journal, 27(3), 379–423.

Sharma, K., Mangaroska, K., van Berkel, N., Giannakos, M., & Kostakos, V. (2021*). Information flow and cognition affect each other: Evidence from digital learning*. International Journal of Human-Computer Studies, 146, 102549.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). *Integrated information theory: From consciousness to its physical substrate*. Nature Reviews Neuroscience, 17(7), 450–461.

Vanmarcke, E. (2010). *Random fields: Analysis and synthesis*. World Scientific.

Vityaev, E. E., Perlovsky, L. I., Kovalerchuk, B. Y., & Speransky, S. O. (2013*). Probabilistic dynamic logic of cognition*.

Carey, S., & Carey, S. (2011). *The Origin of Concepts*. New York: Oxford University Press.

Gómez-Ramirez, J. (2014). *A New Foundation for Representation in Cognitive and Brain Science*. 7, Springer Series in Cognitive and Neural Systems.

Reviewer 1. Andreas Falck, Department of Special Needs Education, University of Oslo

Review of Heidari Beni et al.: "Category theory: From semantic information in the mind to Shannon information in the brain" for SweCog 2023 The authors propose an interesting approach to the mind-body problem, by positing that the mind and the brain can be understood in terms of different types of information. Information in the mind is best described in the everyday sense of meanings. In contrast, neural activity is more readily described in terms of Shannon information, i.e. relations between random variables. Furthermore, the authors claim that the mathematical field of category theory can be recruited to help map these types of information to each other, thus bridging the mind-brain gap. The approach seems interesting and well worth exploring, although it is not clear to me how the mind-body gap is actually bridged. One side of the problem, according to the authors, is to map (Shannon) information in the brain to task goals, which I in turn understand as essentially given in the everyday description of information-as-meaning. In my understanding, the question thus becomes how to distinguish which neural activities or patterns during the performing of a cognitive task that are causally related to that particular task, and which are not – in lay terms, which neural activity is "meaningful" in the context of the task. What I do not understand from the text is how this part of the problem can be solved. The authors describes how tasks and concepts are compositional and can be mathematically related by functors, then they write "Given that a functor associates each task to a concept, we can relate a concept to a random field over neural activities resulting from the equivalent optimization process characterized by the corresponding task, as earlier mentioned." It is not currently clear to me how this optimization process is carried out on the neural level, and thus how the gap is actually bridged. I might also have misunderstood what the paper promises, but some clarification could help either way. Finally, while I find the introduction to the different types of information sufficiently thorough, would recommend a bit more introduction to category theory. As someone hearing about this theory for the first time, I would have appreciated a little more context. (This might in turn help explaining how the gap is bridged).

Reviewer 2. Andreas Chatzopoulos, Dept. of Applied IT, Division of Cognition & Communication, University of Gothenburg

Category theory: From Semantic Information in the mind to Shannon Information in the brain

The paper explores the concept of information as a key factor in bridging the gap between mind and brain. It contrasts the way information is understood at the brain level (neural activities as statistical patterns) with the mind level (representational structures for concepts and meanings). The challenge lies in finding a coherent mathematical relationship between these two forms of information.

Category theory, a mathematical field focused on structures, is proposed as a means to address this challenge of unifying the different levels. Category theory's perspective on optimization is applied to cognitive tasks, relating them to patterns of neural activity. This leads to the idea that each concept or meaning in the mind's representation is associated with structures of information flow in the brain, as defined by Shannon's information theory.

Thus; the text presents possible framework to relate information flows between the brain's neural activities and the mind's representational structures through the application of category theory.

This proposed use of category theory to link information between the brain and the mind is an interesting approach and forms the basis for an intriguing conversation that encompasses the fields of neuroscience, philosophy, and mathematics. Thus, it is my view that this would be a suitable topic for a Cognitive Science conference and constitutes a rich ground for further discussion and elaboration.

Interpreting mental rotation performance in self-described aphantasia through cognitive penetrability

Dániel Pénzes

Author's affiliation: Umeå Universitet

Presenter's e-mail address: daniel.penzes@gmail.com

The depictive view in the mental imagery debate states that mental images have similar spatial structures as their corresponding external object (Kosslyn et al., 1979), meaning that such mental representations resemble or correspond to the external world (Brann, 1991/2017). The everyday expression "the mind's eye" reflects this tradition by referring to perceptual qualities of a mental image. The propositional view, however, contends that beliefs about the external world influence a mental image (the tacit knowledge argument, in Pylyshyn, 1978). This is also known as the cognitive penetrability hypothesis (Pylyshyn, 1979), meaning that higher-level cognitive states "penetrate" or influence a mental image, therefore mental images cannot be described purely in visual terms. Therefore, what the format of a mental image might be becomes a key question in cognitive science (Pearson & Kosslyn, 2015).

People with self-described aphantasia offer a new opportunity to approach this issue. According to Zeman et al. (2015), the term *aphantasia* is coined for those who lack voluntary visual mental images. Traditionally, visual mental imagery is measured by self-reports, such as the Vividness of Visual Imagery Questionnaire (VVIQ, Marks, 1973) or the Object-Spatial Imagery Questionnaire (OSIQ, Blajenkova et al., 2006). The general finding with respect to people with aphantasia is a consistent low average score on the VVIQ as opposed to people without aphantasia (e.g., Milton et al., 2021; Zeman et al., 2015), and a low average score on the object items, as opposed to the spatial items, on the OSIQ (e.g., Dawes et al. 2020; Keogh & Pearson, 2018).

The current study employed the mental rotation task (MRT), where a three-dimensional object needs to be mentally aligned with another one that is rotated to a different angular position (Shepard & Metzler, 1971). This paradigm is an objective way to measure spatial mental imagery performance (Pearson et al., 2013). To test the effects of beliefs on mental images in self-described aphantasia, different instruction conditions were used on the MRT, which was a direct replication of the first study of Borst et al. (2011). The authors first presented one of the three instructions to their participants, after which they performed the MRT, ending the experiment with questionnaires on mental imagery. Two hypotheses were made: if mental rotation is cognitively penetrable, this by itself will not reveal the form of a mental image. If, on the other hand, mental rotation is not cognitively penetrable, this could indicate one possible format: the depictive. Their results showed that irrespective of instruction type, participants' performance on the mental rotation task followed the usual linear trend (increasing angular disparity resulting in increased reaction time), concluding that mental rotation is not cognitively penetrable.

It is yet unknown whether the same applies to aphantasia. Therefore, the following instruction types were acquired directly from Borst et al. (2011). The analytic condition emphasized a piecemeal strategy, explicitly stating not to use any mental rotation. The leap condition referred to a quick mental transformation of the objects, without any further intermediate steps of a rotation. The mental rotation condition used the traditional instruction, to mentally rotate the objects in "the mind's eye" until they are aligned. Twenty-seven participants (21 females, mean age 47 years) with self-described aphantasia completed an online experiment, consisting of the MRT (9 participants in each condition), the OSIQ, and the VVIQ. Participants were recruited from two Facebook groups and provided written consent.

As predicted, participants scored low on the VVIQ (*Mdn* = 16, the lowest possible score) and higher on the spatial items (*Mdn* = 37) than on the object items of the OSIQ (*Mdn* = 19). The analysis of reaction times on the MRT followed the same steps as it was reported by Borst et al. (2011), apart from including non-matching shapes as well. That is, the two objects presented in the MRT can either match (same stimuli) or not (different stimuli). This is in accordance with the guidelines of Metzler and Shepard (1974). First, individual analyses of each condition showed that reaction times increased significantly with increasing angular disparity, both for matching and non-matching shapes, although the analytic condition for non-matching shapes yielded a *p* value of exactly the threshold of alpha level .05. See also figures 1 and 2. Next, correlations between reaction times and angular disparity showed some inconsistencies for both matching and non-matching shapes. That is, not all correlations were statistically significant as it was reported by Borst et al. (2011). The main analysis, however, showed that increasing reaction time with increasing angular disparity was not influenced by different instruction conditions (no interaction effect was found). Additionally, error rates, the speed of processing (in terms of the slopes of the best fitting lines), and the heights of the intercepts (reaction time at 0° angular disparity) were comparable across the conditions, for both matching and non-matching shapes separately. These findings were in line with Borst et al. (2011).

While Borst et al. (2011) did not make a distinction between good and bad mental imagery performance, the current findings replicate their study, with the extension of investigating self-described aphantasia. These results indicate that the theory of cognitive penetrability (Pylyshyn, 1979) is not applicable on mental rotation in aphantasia. This suggests that the depictive view has a stronger stand in the mental imagery debate as opposed to the propositional view, but it also implies that the term *aphantasia* cannot mean a complete lack of mental imagery. Further limitations (e.g., possible repetition effects in the MRT) and future directions (e.g., including additional cognitive processes, such as working memory) are discussed. Given that mental rotation performance in aphantasia is not cognitively penetrable, more research is required to finetune what exactly a lack of mental imagery might mean.

Figure 1

Mean reaction times for different angular disparities of matching stimuli, depending on instruction types.

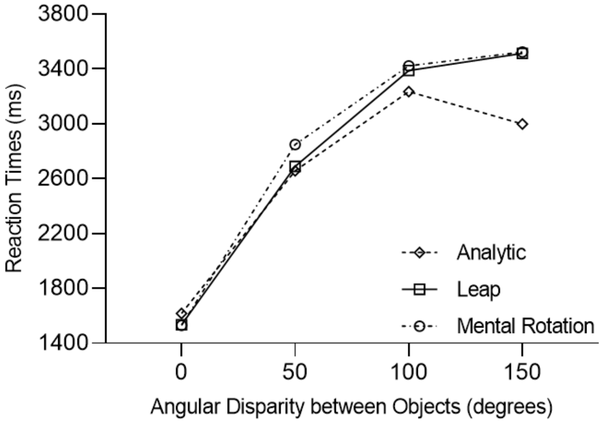
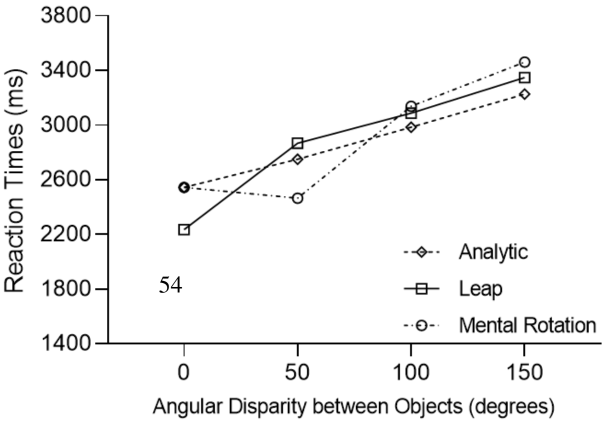


Figure 2

Mean reaction times for different angular disparities of non-matching stimuli, depending on instruction types.



References

Blajenkova, O., Kozhevnikov, M., & Motes, M. A. (2006). Object-spatial imagery: A new self-report imagery questionnaire. *Applied Cognitive Psychology, 20*, 239-263. https://doi.org/10.1002/acp.1182

Borst, G., Kievit, R. A., Thompson, W. L., & Kosslyn, S. M. (2011). Mental rotation is not easily cognitively penetrable. *Journal of Cognitive Psychology*, *23*(1), 60-75. https://doi.org/10.1080/20445911.2011.454498

Brann, E. T. H. (2017). *The world of the imagination: Sum and substance* (25th anniversary ed.). Rowman & Littlefield Publishers (Original work published 1991).

Dawes, A. J., Keogh, R., Andrillon, T., & Pearson, J. (2020). A cognitive profile of multi-sensory imagery, memory and dreaming in aphantasia. *Scientific Reports*, *10*, Article 10022. https://doi.org/10.1038/s41598-020-65705-7

Keogh, R., & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex*, *105*, 53-60. https://doi.org/10.1016/j.cortex.2017.10.012

Kosslyn, S. M., Pinker, S., Smith, G. E., & Shwartz, S. P. (1979). On the demystification of mental imagery. *The Behavioral and Brain Sciences*, *2*(4), 535-581. https://doi.org/10.1017/S0140525X00064268

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, *64*, 17-24. https://doi.org/10.1111/j.2044-8295.1973.tb01322.x

Metzler, J., & Shepard, R. N. (1974). Transformational studies of the internal representation of three-dimensional objects. In R. L. Solso (Ed.), *Theories in Cognitive Psychology: The Loyola Symposium*. Lawrence Erlbaum.

Milton, F., Fulford, J., Dance, C., Gaddum, J., Heuerman-Williamson, B., Jones, K., Knight, K. F., MacKisack, M., Winlove, C., & Zeman, A. (2021). Behavioral and neural signatures of visual imagery vividness extremes: Aphantasia versus hyperphantasia. *Cerebral Cortex Communications*, *2*(2), 1-15. https://doi.org/10.1093/texcom/tgab035

Pearson, D. G., Deeprose, C., Wallace-Hadrill, S. M. A., Burnett Heyes, S., & Holmes, E. A. (2013). Assessing mental imagery in clinical psychology: A review of imagery measures and a guiding framework. *Clinical Psychology Review*, *33*(1), 1-23. https://doi.org/10.1016/j.cpr.2012.09.001

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. PNAS, 112(33). 10089-10092. https://doi.org/10.1073/pnas.1504933112

Pylyshyn, Z. W. (1978). *Imagery and Artificial Intelligence*. University of Minnesota Press, Minneapolis. https://hdl.handle.net/11299/185336

Pylyshyn, Z. W. (1979). The rate of "mental rotation" of images: A test of a holistic analogue hypothesis. *Memory & Cognition*, *7*(1), 19-28. https://doi.org/10.3758/BF03196930

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703. https://doi.org/10.1126/science.171.3972.701

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery – Congenital aphantasia. *Cortex*, *73*, 378-380. https://doi.org/10.1016/j.cortex.2015.05.019

Reviewer 1. Valentina Fantasia, Department of Philosophy, Lund University

Comment(s): The article is clear, concise and well-written. The topic is interesting and can certainly positively contribute to the conference. However, in my opinion, a few things need to be clarified and/or addressed in the work: What was the aim of the study, and how is the study expected to contribute to the current debate or add to pre-existing literature on the subject? Participants: a) were recruited on two fb groups: a) Were there any selection criteria or no selection at all? and if so, which ones? These aspects are all relevant since aphantasia is described as lacking voluntary visual mental images. So recruitment process and criteria may well be relevant (and potentially biased) in regard to how people perceive themselves about their mastering of visual mental imaginary capacity. ; b) did you counterbalance participants in the different conditions? c) did you check if there is any effect of age or sex? Some examples of how we use/apply visual mental images as mental rotation in our everyday life would be helpful. What are the implications of your results in the daily ecology of people's life? I think your conclusions are interesting but need to be more cautious and less generalised. Maybe drawing connections with/ touching on other possible competing or concurrent cognitive factors as influencing variables?

Reviewer 2. Daniel Sjölie, Division of Informatics, University West

The consideration of how previous results on mental rotation might be affected by aphantasia, and the corresponding connection to the format of mental images and imagery, is intriguing and the presented results are interesting. The experimental setup, with mental rotation under different instructions, provides a clear example of a situation where higher-level beliefs can be hypothesized to affect lower-level cognition. However, to better adapt for a diverse cognitive science readership and highlight the novel contribution, the clarity of the paper could be improved. As a specific example, symbolic representations commonly refer to a setup where it is only the abstract identity of the symbol that matters for what it represents, not its spatial structure or similar. The current text about depictive mental images seems to fit poorly with this, and the assertion that a representation is symbolic because it symbolizes something in the external world (irrespective of how?) is potentially confusing. If there is space to expand the text, or in an associated presentation, it may be good to take another look at how such concepts are introduced. It would also be good to have more clarity concerning how the present study relates to Borst et al. (2011). It is mentioned that the study is a direct replication, but what exactly Borst et al. did, and what the present study adds to this replication, is sprinkled out throughout the text. I do not doubt that this can be satisfactorily answered, but a little more structure would help. Finally, the next to last sentence is confusing, highlighting the importance of beliefs while the results presented here do not support the importance of beliefs in this situation, as in believing that you cannot see a mental image rotate does not change your performance in mental rotation.

Author response from Pénzes

Review 1, By Valentina Fantasia

Comment(s):
The article is clear, concise and well-written. The topic is interesting and can certainly positively contribute to the conference. However, in my opinion, a few things need to be clarified and/or addressed in the work:
What was the aim of the study, and how is the study expected to contribute to the current debate or add to pre-existing literature on the subject?

- **The aim of the study was to reexamine the possibility of aphantasia with a mental imagery definition (i.e., the propositional view) that is not typically used when people talk about lacking "visual" mental images. Hence the title of my study: "Interpreting …". If there are no visual qualities to their mental images, could it perhaps be propositional? Perhaps, cognition penetrates mental imagery (Pylyshyn, 1979), therefore not being able to describe the experience with terms such as "vividness" or "seeing with the mind's eye"?**
- **My results do not seem to support this idea within one narrow field – the mental rotation task – yet it aims at offering a new perspective on understanding what a lack of mental imagery might mean. The next step would be to expand the scope of experimental paradigms on aphantasia, for example, by finding ways to implement the theory of cognitive penetrability on other experimental designs.**

Participants: a) were recruited on two fb groups: a) Were there any selection criteria or no selection at all? and if so, which ones? These aspects are all relevant since aphantasia is described as lacking voluntary visual mental images. So recruitment process and criteria may well be relevant (and potentially biased) in regard to how people perceive themselves about their mastering of visual mental imaginary capacity. ;

- **The only selection criterion I used was that I was calling for participants who claimed to have no visual mental imagery.**
- **Your feedback raises an interesting question that I touched upon only briefly in my study. The original study that coined the term 'aphantasia' (Zeman et al., 2015) was based on 21 volunteering participants, people who directly contacted the authors. One tradition in mental imagery research is to verify "vividness" of mental images through self-reports, for example the Vividness of Visual Imagery Questionnaire (VVIQ, Marks, 1973), which was also used by Zeman et al. (2015). An overlooked detail about this study can be read from their supplementary document; the email sent to these voluntary participants states that "We expect you to rate your imagery at the low end of the spectrum …" (p1.). That is, a potential bias emerges here.**
- **Thus, on the one hand I aimed at following the same procedure of recruitment (without emphasizing my expectations): asking for voluntary participants and providing them the VVIQ. On the other hand, I am aware of a potential bias here that invites a path for future investigations.**

b) did you counterbalance participants in the different conditions?

- **Nine participants were randomly allocated to each condition in the mental rotation task (27 in total). The four orientations (0°, 50°, 100°, & 150°) were presented randomly but equally often, and half of the trials contained matching shapes.**
- **No additional counterbalancing was employed. There was no need to assign participants to different conditions where the order of the tasks would change,**

since all participants had to first receive the instruction, right after which the mental rotation task occurs. The two questionnaires were placed at the end, so that participants would not think consciously about their mental imagery experiences before performing the MRT.

c) did you check if there is any effect of age or sex?

- **As the sample was skewed towards more females than males (21 and six, respectively) with a relatively high mean age (47 years), the effects of age or sex were not checked. I did, however, mention possible sex differences in the Discussion part of my study. Traditional literature on the mental rotation task shows that men outperform women (e.g., Voyer et al., 1995), while Scheer et al. (2018) found no evidence for such sex differences.**

Some examples of how we use/apply visual mental images as mental rotation in our everyday life would be helpful.

- **A common scenario would be when we go to Ikea. Upon seeing the desired table, we might try to visualize our dining room, how the table would fit a specific corner, or perhaps that we need to rotate it so that it fits with the rest of the furniture.**
- **Moen et al. (2020) brings up many good real-life examples, such as when we are reading a map. We might try to visualize the buildings, the directions, and our body's position in the landscape. Or perhaps when we fill up a dishwasher, and we judge whether that big plate fits the upper part of the machine, next to the glasses.**

What are the implications of your results in the daily ecology of people's life?

- **Given that we accept that mental rotation performance is not cognitively penetrable in self-described aphantasia, this points to one possible conclusion: aphantasia still retains some sort of spatial visual mental imagery. That is, the label "aphantasia" cannot fully cover what people might (or might not) experience. A problem in interpretation may arise when people use this label.**
- **One interesting point that perhaps connects to your question is the following. The two Facebook groups (for people with aphantasia) contain an abundance of comments and discussions on how people experience their aphantasia and how they live their lives with it. A whole community emerged online. One such theme is when people explain their obstacles in life (such as those appearing in school or at work) in terms of how aphantasia is the reason for them. This could certainly be an interesting topic for a future qualitative study, for example.**

I think your conclusions are interesting but need to be more cautious and less generalised. Maybe drawing connections with/ touching on other possible competing or concurrent cognitive factors as influencing variables?

- **I rephrased some parts of my conclusions as the second review pointed out another issue as well.**
- **In my study I touch upon one particular cognitive factor: working memory. Studies such as Jacobs et al. (2018) or Monzel et al. (2022) showed that visual working memory (but also verbal memory, short- and long-term) performance was impaired in people with aphantasia, compared to people without aphantasia. It can certainly be argued that performance on the mental rotation task is a**

demanding or challenging (or perhaps even stressful) task, since participants had a fixed time limit (quite short) to respond in each trial.

Review 2, By Daniel Sjölie

The consideration of how previous results on mental rotation might be affected by aphantasia, and the corresponding connection to the format of mental images and imagery, is intriguing and the presented results are interesting. The experimental setup, with mental rotation under different instructions, provides a clear example of a situation where higher-level beliefs can be hypothesized to affect lower-level cognition. However, to better adapt for a diverse cognitive science readership and highlight the novel contribution, the clarity of the paper could be improved.

As a specific example, symbolic representations commonly refer to a setup where it is only the abstract identity of the symbol that matters for what it represents, not its spatial structure or similar. The current text about depictive mental images seems to fit poorly with this, and the assertion that a representation is symbolic because it symbolizes something in the external world (irrespective of how?) is potentially confusing. If there is space to expand the text, or in an associated presentation, it may be good to take another look at how such concepts are introduced.

- **Brann in her book (1991/2017) contrasts a more traditional – philosophical – account of a mental representation, emphasizing correspondence and resemblance between the external and the mental world, to the computational definition of classical cognitive science, that is, symbols being operated by rules in a formal system. Unfortunately, I did not notice at first that I mixed these two notions together; I corrected this in the short paper.**
- **As a matter of fact, the definition of a mental representation is not strictly relevant to my study. I touch upon this only briefly so that the connection between my subject matter to cognitive science becomes apparent.**

It would also be good to have more clarity concerning how the present study relates to Borst et al. (2011). It is mentioned that the study is a direct replication, but what exactly Borst et al. did, and what the present study adds to this replication, is sprinkled out throughout the text. I do not doubt that this can be satisfactorily answered, but a little more structure would help.

- **I added more details about this study to my short paper.**

Finally, the next to last sentence is confusing, highlighting the importance of beliefs while the results presented here do not support the importance of beliefs in this situation, as in believing that you cannot see a mental image rotate does not change your performance in mental rotation.

- **Thank you for pointing this out. I realize that I confounded two separate concepts (the idea of self-describing with something and the idea of the propositional definition of a mental image) without explicitly stating how the two may or may not connect. I abandoned this topic for now in my revised short paper, as it shows rather a personal intuition irrelevant to the subject matter.**

References

Brann, E. T. H. (2017). *The world of the imagination: Sum and substance* (25th anniversary ed.). Rowman & Littlefield Publishers (Original work published 1991).

Borst, G., Kievit, R. A., Thompson, W. L., & Kosslyn, S. M. (2011). Mental rotation is not easily cognitively penetrable. *Journal of Cognitive Psychology*, *23*(1), 60-75. https://doi.org/10.1080/20445911.2011.454498

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, *64*, 17-24. https://doi.org/10.1111/j.2044-8295.1973.tb01322.x

Moen, K. C., Beck, M. R., Saltzmann, S. M., Cowan, T. M., Burleigh, L. M., Butler, L. G., Ramanujam, J., Cohen, A. S., & Greening, S. G. (2020). Strengthening spatial reasoning: Elucidating the attentional and neural mechanisms associated with mental rotation skill development. *Cognitive Research: Principles and Implications*, *5*(1), 20. https://doi.org/10.1186/s41235-020-00211-y

Pylyshyn, Z. W. (1979). The rate of “mental rotation” of images: A test of a holistic analogue hypothesis. *Memory & Cognition*, *7*(1), 19-28. https://doi.org/10.3758/BF03196930

Scheer, C., Mattioni Maturana, F., & Jansen, P. (2018). Sex differences in a chronometric mental rotation test with cube figures: A behavioral, electroencephalography, and eye-tracking pilot study. *Neuroreport*, *29*(10), 870-875. https://doi.org/10.1097/WNR.0000000000001046

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*(2). 250-270. https://doi.org/10.1037/0033-2909.117.2.250

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery – Congenital aphantasia. *Cortex*, *73*, 378-380. https://doi.org/10.1016/j.cortex.2015.05.019

An attempt to replicate the Response Conflict/Competition Theory's hypothesis of the Stroop effect with functional near infrared spectroscopy

Simon Skau[1], Lina Bunketorp-Käll[1], Helena Filipsson Nyström [2], Hans-Georg Kuhn[1]

[1]*Institute of Neuroscience and Physiology, Department of Clinical Neuroscience, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.*

[2]*Department of Endocrinology, Sahlgrenska University Hospital, Gothenburg, Sweden*

simon.skau@gu.se

Introduction

In the Stroop task, which is a conflict processing task, participants are presented with color words (e.g., RED or BLUE) written in ink colors that may or may not match the word's meaning (e.g., red or blue). The participants are instructed to respond to the ink color, which is the task-relevant stimulus dimension, while ignoring the semantic meaning of the word, which is the task-irrelevant stimulus dimension. The Response Conflict/Competition Theory (RCCT) suggests that the Stroop effect occurs because the brain simultaneously processes both the task-relevant and task-irrelevant stimulus dimensions, activating the associated response tendencies. Specifically, in congruent trials, both dimensions elicit the same response, while in incongruent trials, they elicit conflicting responses. According to the RCCT, the longer response times observed in incongruent trials are a result of suppressing the incorrect response tendency (Szűcs, Killikelly, & Cutini, 2012). Therefore, there should be some preparatory activation detectable in the brain for a starting of an incorrect response during incongruent trials but not for congruent trials (Figure 1A and 1B for illustration).

To investigate the RCCT, Szücs et al. conducted a series of studies utilizing electroencephalography (EEG) during various versions of the Stroop task (Bryce, Szűcs, Soltész, & Whitebread, 2011; Szűcs & Soltész, 2007, 2008; Szűcs, Soltész, Bryce, & Whitebread, 2009; Szűcs, Soltesz, Jarmi, & Csepe, 2007; Szűcs, Soltész, & White, 2009). They focused on measuring activity over the motor cortices to identify preparatory response activity in the ipsilateral motor cortex during incongruent trials. Two of their studies reported evidence of preparatory response activity lasting 100 ms in the ipsilateral motor cortex (Bryce et al., 2011; Szűcs et al., 2009). However, in their other studies, the absence of such activity was attributed to the measurement being influenced by electrical activity in both motor cortices (Szűcs et al., 2012). To address this limitation, Szücs et al. subsequently conducted a follow-up study using functional near-infrared spectroscopy (fNIRS), an optical imaging technique that measures changes in oxygenated (oxy-Hb) and deoxygenated hemoglobin concentrations as an indirect measure of cortical activation. With its spatial resolution of 1 cm^3, fNIRS was expected to be specific enough to avoid interference from the other hemisphere, and its temporal resolution of 10 Hz was sufficient to detect preparatory activity. Consistent with the RCCT, the authors hypothesized and found evidence of a larger increase in oxy-Hb in the ipsilateral cortex to the response hand during the ascending phase (2.5-4.5 seconds) of the hemodynamic response in incongruent trials compared to congruent trials. This finding supported the RCCT and demonstrated the suitability of fNIRS for detecting preparatory activity by examining transient increases in the hemodynamic response (Szűcs et al., 2012). However, it is important to note that their study had a relatively small sample size, consisting of only twelve participants, highlighting the need for replication. Therefore, the present study aims to evaluate the RCCT in a larger cohort using fNIRS imaging in an Animal Stroop test, as previously done by (Szűcs et al., 2012).

Method

The current paper is based on data collected on control subjects from the CogThy project (Holmberg et al., 2019). The fNIRS part of the study was started in May of 2015. The Regional Ethical Review Board in Gothenburg approved the current study (reference number: T955-14/190-10). Twenty-eighth women with a mean age of 35.3 ± 9.9 years were recruited to the current study.

The experiment comprised three parts: a pretest, an intermediate task, and a posttest. The pre- and posttest utilized the Animal Stroop test, which lasted for 20 minutes, while the intermediate task involved a 30-minute reading comprehension test from the Swedish Scholastic Aptitude Test (SweSAT). The Animal Stroop test employed a size congruency task (Szűcs et al., 2012), where participants were simultaneously presented with images of two different animals—one on the left side and one on the right side of the screen (see Figure 1B).

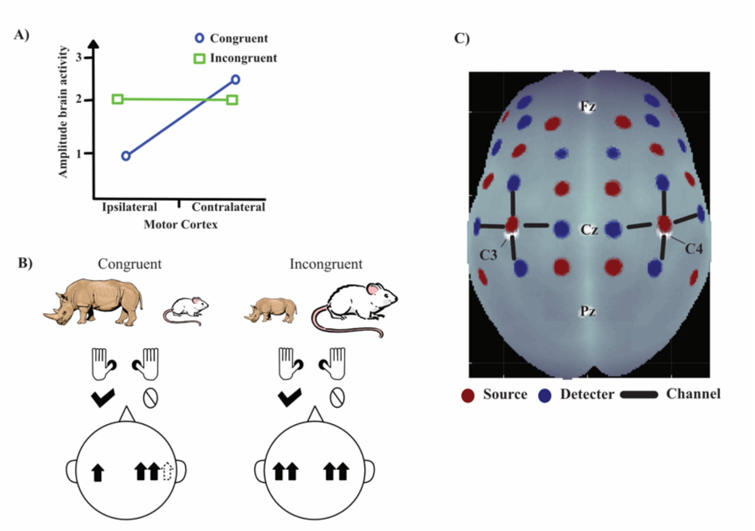


Figure 1. Stroop effect and measurement. A) Predicted activity over the motor cortices based on the RCCT. B) Example of the congruent and incongruent trials in the Animal Stroop task. The upwards arrows illustrate the same type of activity as in A), where the two arrows mean higher amplitude of brain activity. The dashed arrow indicates that the RCCT hypotheses activity Congruent ≥ Incongruent in contralateral cortex. C) Visualization of the fNIRS measurements from a transverse view. Red dots are light sources, blue dots are detectors, black lines are estimated channels, white dots are 10/20 landmarks.

Their objective was to determine which of the two animals was larger in real life, irrespective of their size on the screen. Participants indicated their response by pressing either their left or right thumb on a gamepad. In congruent trials, the larger animal appeared as the bigger image, while in incongruent trials, the larger animal appeared as the smaller image. Each test comprised 80 trials, semi-randomized to generate 20 trials for congruent left-hand response, congruent right-hand response, incongruent left-hand response, and incongruent right-hand response. The response to stimulus interval ranged from 10 to 14 seconds. The task was divided into four blocks with a 30-second break between each block.

To replicate the study conducted by Szücs et al. (2012) we used a total of 160 trials by combining the pre- and posttest tasks. For fNIRS analysis, we utilized the statistical package in MATLAB 2018b (for code and functions used, https://github.com/SimonSkau/RCCT_replication). Following the approach of Szücs et al., we performed a repeated measures ANOVA with the within-factors of Congruency (congruent, incongruent) and Lateralization (ipsilateral, contralateral), as well as the interaction of Congruency vs. Lateralization, for each timepoint from second 0 to 12 (yielding a total of 120 timepoints) for oxy-Hb. Consistent with Szücs et al. (2012), a period of ten or more consecutive timepoints with a p-value < .05 was considered significant. If a significant period was identified, we calculated the average concentration during that period and further conducted paired t-tests to examine main effects. Only channels over the moter cortex are anlysised here (Figure 1C).

Result

Mean oxy-Hb responses are visualized in Figure 2A-B). For the point-by-point repeated ANOVA there were no significant period for *Congruency* (congruent, incongruent) with only nine significant timepoints in a row between 1.8 to 2.6 seconds after stimulus onset. For Lateralization (ipsilateral, contralateral) there were a significant period, with 31 significant timepoints in a row between 4.5 and 7.7 seconds after stimulus onset. Post hoc test showed higher oxy-Hb in contralateral motor cortex with t(21)=2.8768, p = 0.009 CI [0.0210 0.1305] with a Cohens d of 0.61 (Figure 2D). There was no significant interaction (*Congruency* vs. *Hemisphere*), failing to replicate the result from Szücs et al. (2012) and the RCCT.

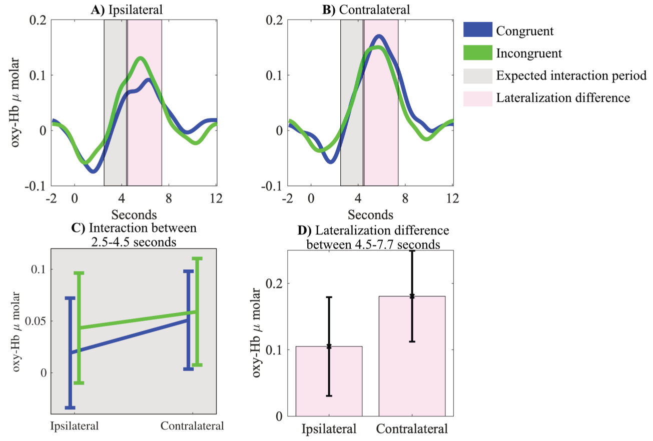


Figure 2. A) and B) are average oxy-Hb response for ipsilateral and contralateral primary motor cortex respectively. Congruent in blue and incongruent in green. Gray area indicates the expected interaction period from Szücs *et al.* and pink area is the significant period for Lateralization. C) Show the result in the expected interaction period, should be compared too Figure 1A. D) shows the difference in lateralization period. Error bars are standard deviation of the mean.

Discussion

The Response Conflict/Competition Theory (RCCT) posits that the incongruent condition of a Stroop task elicits different responses for the task-relevant and -irrelevant stimulus dimension, whereas in the congruent condition, similar responses are expected. According to the RCCT, parallel processing in the brain during incongruent trials should generate preparatory activity, while such activity is not anticipated during congruent trials. In the context of the Animal Stroop task, preparatory activity is hypothesized to be present in the primary motor cortex ipsilateral to the response hand during incongruent trials, and higher activity in the contralateral cortex during congruent trials. Previous research by Szücs et al. (2012) demonstrated an interaction between Congruency (congruent, incongruent) and Lateralization (ipsilateral, contralateral) in the oxy-Hb levels during the time interval of 2.5 to 4.5 seconds. They reported a greater increase in oxy-Hb for incongruent trials than for congruent trials in the ipsilateral motor cortex, along with higher activity in the contralateral cortex during congruent trials, aligning with the predictions of the RCCT.

In our study, however, we were unable to replicate these findings (Figure 2C, compared to the predicted result in Figure 1A). Instead, we observed a Lateralization effect around the peak (Figure 2A-B and D), indicating overall higher activity in the contralateral motor cortex, as expected. In conclusion, our results suggest that both incongruent and congruent trials elicit greater activity in the contralateral motor cortex compared to the ipsilateral cortex. If the RCCT's preparatory signal exists, it appears to be very brief, and larger sample sizes would be necessary to detect it effectively. However, another question is whether preparatory activity could be detected in the ascending phase due to the sluggish nature of the hemodynamic response.

References

Bryce, D., Szűcs, D., Soltész, F., & Whitebread, D. (2011). The development of inhibitory control: An averaged and single-trial Lateralized Readiness Potential study. *NeuroImage (Orlando, Fla.), 57*(3), 671-685.

Holmberg, M., et al (2019). Structural brain changes in hyperthyroid Graves' disease: protocol for an ongoing longitudinal, case-controlled study in Göteborg, Sweden—the CogThy project. *BMJ Open, 9*(11).

Szűcs, D., Killikelly, C., & Cutini, S. (2012). Event-related near-infrared spectroscopy detects conflict in the motor cortex in a Stroop task. *Brain Res, 1477*, 27-36.

Szűcs, D., & Soltész, F. (2007). Event-related potentials dissociate facilitation and interference effects in the numerical Stroop paradigm. *Neuropsychologia, 45*(14), 3190-3202.

Szűcs, D., & Soltész, F. (2008). The interaction of task-relevant and task-irrelevant stimulus features in the number/size congruency paradigm: an ERP study. *Brain research, 1190*, 143.

Szűcs, D., Soltész, F., Bryce, D., & Whitebread, D. (2009). Real-time tracking of motor response activation and response competition in a Stroop task in young children: a lateralized readiness potential study. *J Cogn Neurosci, 21*(11), 2195-2206.

Szűcs, D., Soltesz, F., Jarmi, E., & Csepe, V. (2007). The speed of magnitude processing and executive functions in controlled and automatic number comparison in children: an electro-encephalography study. *Behavioral and Brain Functions, 3*(23), 23.

Szűcs, D., Soltész, F., & White, S. (2009). Motor conflict in Stroop tasks: Direct evidence from single-trial electro-myography and electro-encephalography. *NeuroImage (Orlando, Fla.), 47*(4), 1960-1973.

Reviewer 1. Birger Johansson, Department of Philosophy, Lund University

This short paper presents a replication of a Stroop task experiment, as originally conducted by Szűcs et al. (2012). In a Stroop task, reaction times are increased during incongruent trials, and the Response Conflict/Competition Theory (RCCT) posits that these longer reaction times result from the inhibition of incorrect response. This inhibition should be visible as preparatory brain activation.

The original authors (Szűcs et al., 2012) recorded the activity in the motor cortex, using functional near-infrared spectroscopy (fNIRS), to identify any preparatory response activity and found higher activation in the ipsilateral cortex to the response hand in incongruent trials compared to the congruent trials.

This short paper replicates the experiment, with an increased number of participants, but finds no increased activation as reported in the original study. The paper is well-written, easy to read and the authors describe the Stroop effect and the Response Conflict/Competition Theory in a clear way

The authors heavily rely on the 2012 experiment by Szűcs et al., which is critical in replication studies. However, there are some uncertainties regarding the similarity between this replication and the original paper. For instance, the authors use different terminologies: 'Task-relevant stimulus dimension' versus 'stimulus-relevant dimension'. Is there a specific reason for this? If so, this can be explained to the reader. How similar is the experiment to the original experiment? Are the authors using exactly the same pictures as in the original experiment? Are the stimuli presented in a similar way as in the original experiment? Size of the screen presenting the stimuli?

The paper has very limited references other than those to Szűcs. Not even Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18(6), 643–662 is referenced.

There is no reference to figure 1C in the text. If figure 1C is an important part of the experiment, the figure needs to be explained in the text.

In contrast to the original study's dual objectives, the short paper does not address the suitability of fNIRS for detecting motor cortex preactivity. The conclusion is that "If the RCCT's preparatory signal exists, it appears to be very brief, and larger sample sizes would be necessary to detect it effectively". Could it be the case that this method is not suited for the task?

The authors have made an important work in replicating a scientific experiment and I appreciate reading the short paper.

Reviewer 2. Linus Holm, Department of Psychology, Umeå University

The paper attempts to replicate a study by Szücs and colleagues (2012). In the original paper, the hemodynamic response assessed via functional near infra-red spectroscopy (fNIRS) suggested an interaction between stroop task congruency and hemisphere such that incongruent tasks produced a stronger ipsilateral response than congruent tasks. The effect was taken to imply response competition prior to action and support the Response Conflict/Competition Theory (RCCT). Skau and colleagues did not replicate this main result despite adopting the same stimulus material, task, and testing twice as many participants. It is good to see attempts at replication now and then, and the paper by Skau and colleagues constitutes a solid example. It does however mean that shortcomings in the original study may propagate over to the replication. Please consider that my criticism then might be attributed to the original study as you react to my review. Also, I am no expert in fNIRS and some comments relating to the method may therefore be naïve – please just bear with me.

1. (How) is it possible to identify the preparatory brain signals via assessment of the hemodynamic response in the study? It seems like Szücs (2012) essentially determine period by splitting the aggregated hemodynamic response signal into an ascending and a descending part based on peak response. I may very well have misunderstood this but how can one map the sluggish hemodynamic response, which propagates over several seconds, onto preparation and action, respectively, when the task itself presumably only lasts a second or so?
2. Following up on this, if mapping to task period is challenging to achieve, could the theory be saved by considering interactions during the latter, or indeed the entire hemodynamic response period? I presume post-choice activity could potentially also interact in the same way as preparation, and then reflect e.g., reflection (residual activity – I understand the target locus is motor cortex here) on the action but that might be a question for later studies to distinguish between.
3. Continuing this reasoning (but also as an independent point) – could you please report the full ANOVA outcome also for the "lateralization period"? As of now, you only state in text that the interaction was not statistically reliable, but looking at panels 2A and 2B it looks like there is some support for an interaction. It would be good to see the corresponding interaction in 2D as is currently portrayed in 2C.
4. My interpretation is that the authors suggest the null effect to reflect a potential absence of effect. The analysis would then improve by testing this directly using Bayesian statistics instead and report the likelihood for the null hypothesis vs the alternative.
5. The introduction states that incorrect incongruent trials should differ from incorrect congruent trials. To me this is a bit unclear. On (I suspect the rare) event that incorrect responses are made in the congruent condition, is it not possible to think that they might reflect a conflict for an inappropriate response too? As of now (but very possibly due to me misreading the methods and results) – it seems the conflict might be a result also of performance – it would then follow that there should have been less conflict for correct than incorrect trials, partly independent of congruence.

Author response from Skau et al.

Reply to all reviewers:

We would like to thank all three reviewers for their work and valuable feedback.

Review 1

By Birger Johansson

This short paper presents a replication of a Stroop task experiment, as originally conducted by Szűcs et al. (2012). In a Stroop task, reaction times are increased during incongruent trials, and the Response Conflict/Competition Theory (RCCT) posits that these longer reaction times result from the inhibition of incorrect response. This inhibition should be visible as preparatory brain activation.

The original authors (Szűcs et al., 2012) recorded the activity in the motor cortex, using functional near-infrared spectroscopy (fNIRS), to identify any preparatory response activity and found higher activation in the ipsilateral cortex to the response hand in incongruent trials compared to the congruent trials.

This short paper replicates the experiment, with an increased number of participants, but finds no increased activation as reported in the original study. The paper is well-written, easy to read and the authors describe the Stroop effect and the Response Conflict/Competition Theory in a clear way

The authors heavily rely on the 2012 experiment by Szűcs et al., which is critical in replication studies. However, there are some uncertainties regarding the similarity between this replication and the original paper. For instance, the authors use different terminologies: 'Task-relevant stimulus dimension' versus 'stimulus-relevant dimension'. Is there a specific reason for this? If so, this can be explained to the reader.

Reply: In order to improve consistency of terminology we have now changed the term to be *task-relevant stimulus dimension*.

How similar is the experiment to the original experiment? Are the authors using exactly the same pictures as in the original experiment?

Reply: The same set of images was used (we received the images from Szűcs).

Are the stimuli presented in a similar way as in the original experiment?

Reply: We designed the trials (pair of animals) and order of the trials independently and they were thus not identical to the original study. We only combined animal pairs with a non-ambiguous difference in real life sizes, i.e., we chose rhino vs mouse, but not dog vs cat.

Size of the screen presenting the stimuli?

Reply: The size of the screen was 24 inches. Participants sat 1m away from it. The larger picture was double the size of the smaller image.

The paper has very limited references other than those to Szűcs. Not even Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18(6), 643–662 is referenced.

Reply: Unfortunately, there is a strict limitation of space for these very short papers. We were therefore forced to limit the references to a minimum and we thus did not include e.g. Stroop (1935).

There is no reference to figure 1C in the text. If figure 1C is an important part of the experiment, the figure needs to be explained in the text.

Reply: We have now referenced figure 1C towards the end of the method section.

In contrast to the original study's dual objectives, the short paper does not address the suitability of fNIRS for detecting motor cortex preactivity. The conclusion is that "If the RCCT's preparatory signal exists, it appears to be very brief, and larger sample sizes would be necessary to detect it effectively". Could it be the case that this method is not suited for the task?

Reply: This is a genuine possibility. This is however not something that could be evaluated within the frame of this study. It should rather be evaluated based on a pure motor task, where one knows that one would detect a lateralize readiness potential (LRP). In other words, it could be the case that:

A) fNIRS cannot detect an LRP, here there are mix results e.g., Drenckhahn et al 2015.

B) this Stroop task cannot generate an LRP,

C) the Stroop task can generate an LRP but it requires a much higher power to be detected

D) Independent from fNIRS data, a "size distance effect" is possible. The distance effect has been shown with numbers, that when deciding which number/amount is higher/larger, processing the size difference requires a longer time the closer two numbers are e.g., 2 vs. 4 compared to 2 vs. 8. Kadosh and colleagues used a number size congruency version of the Number Stroop test and focused on the difference in the activity in the motor cortexes for different distances. They found an LRP when there was a high distance between the numbers (e.g., 2 and 8), but it disappeared when there was a low distance (e.g., 2 and 4) (Kadosh et al. 2007). A similar distance effect is possible for comparisons of animal sizes as well, but harder to control for, since size familiarizations for animal sizes probably vary to a larger extent between individuals than for numbers. This is not likely to have an impact on our study, since we find a lateralization effect, but it could possibly explain why we do not find a congruency effect. However, it is only according to the RCCT that there should be a congruency affect in the motor cortex in this type of test.

The authors have made an important work in replicating a scientific experiment and I appreciate reading the short paper.

Review 2

By Linus Holm

The paper attempts to replicate a study by Szücs and colleagues (2012). In the original paper, the hemodynamic response assessed via functional near infra-red spectroscopy (fNIRS) suggested an interaction between stroop task congruency and hemisphere such that incongruent tasks produced a stronger ipsi-lateral response than congruent tasks. The effect was taken to imply response competition prior to action and support the Response Conflict/Competition Theory (RCCT). Skau and colleagues did not replicate this main result despite adopting the same stimulus material, task, and testing twice as many participants. It is good to see attempts at replication now and then, and the paper by Skau and colleagues constitutes a solid example. It does however mean that shortcomings in the original study may propagate over to the replication. Please consider that my criticism then might be attributed to the original study as you react to my review. Also, I am no expert in fNIRS and some comments relating to the method may therefore be naïve – please just bear with me.

1. *(How) is it possible to identify the preparatory brain signals via assessment of the hemodynamic response in the study? It seems like Szücs (2012) essentially determine period by splitting the aggregated hemodynamic response signal into an ascending and a descending part based on peak response. I may very well have misunderstood this but how can one map the sluggish hemodynamic response, which propagates over several seconds, onto preparation and action, respectively, when the task itself presumably only lasts a second or so?*

Reply: Indeed, what the reviewer has pointed out is a real issue. Another complication is that with an increased neural activity there should be an initial dip in the concentration of oxy-Hb before the vascular response starts. This initial dip happens between 1-2 seconds after stimulus onset. One possibility is to evaluate preparatory activity by measuring the initial dip. However, that timescale is rather short. But an initial dip would affect the ascending phase as well. Usually, the peak of the hemodynamic response is analyzed, in order to handle the problems you raise. Therefore, it might be better in this case restrict analysis to only investigate the peak. However, if the preparatory activity only lasts for 100ms, then it might only show a small effect on the overall hemodynamic response and thus many more trials and participants would be needed in order to detect it.

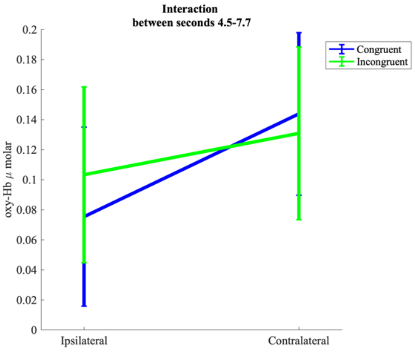
We have now added an additional sentence about this in the discussion.

2. *Following up on this, if mapping to task period is challenging to achieve, could the theory be saved by considering interactions during the latter, or indeed the entire hemodynamic response period? I presume post-choice activity could potentially also interact in the same way as preparation, and then reflect e.g., reflection (residual activity – I understand the target locus is motor cortex here) on the action but that might be a question for later studies to distinguish between.*

Reply: This is a very interesting point. If there is a post-choice activity, however brief, it would propagate together with the pre-choice activity and affect the whole hemodynamic response. However, we assume that it would not affect the descending phase, since post-choice activity is still rather fast, and should thus be involved in generating a higher peak. Again, the peak would be a better way to evaluate this.

3. *Continuing this reasoning (but also as an independent point) – could you please report the full ANOVA outcome also for the "lateralization period"? As of now, you only state in text that the interaction was not statistically reliable, but looking at panels 2A and 2B it looks like there is some support for an interaction. It would be good to see the corresponding interaction in 2D as is currently portrayed in 2C.*

Reply: In the attached figure we visualize the interaction for the time-period in figure 2D (the bar graph) and the following table shows the full ANOVA for that time span. Note that in the paper the post hoc test is done with a paired t-test, thus the numbers reported here are not identical to those reported in the paper. The lateralization effect seems to be driven only by the difference in the congruent condition, and not the incongruent condition, giving some credence to the RCCT. This time interval involves the peak.



Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2_p
Congruency	0.001	1	0.001	0.094	0.763	0.004
Residuals	0.275	21	0.013			
Lateralization	0.051	1	0.051	10.786	0.004	0.339
Residuals	0.099	21	0.005			
Congruency ✻ Lateralization	0.009	1	0.009	1.019	0.324	0.046
Residuals	0.188	21	0.009			

Note. Type III Sum of Squares

4. *My interpretation is that the authors suggest the null effect to reflect a potential absence of effect. The analysis would then improve by testing this directly using Bayesian statistics instead and report the likelihood for the null hypothesis vs the alternative.*

Reply: This is a great point. We will try to do this in upcoming experiments.

5. *The introduction states that incorrect incongruent trials should differ from incorrect congruent trials. To me this is a bit unclear. On (I suspect the rare) event that incorrect responses are made in the congruent condition, is it not possible to think that they might reflect a conflict for an inappropriate response too? As of now (but very possibly due to me misreading the methods and results) – it seems the conflict might be a result also of performance – it would then follow that there should have been less conflict for correct than incorrect trials, partly independent of congruence.*

Reply: We agree that our phrasing may have led to misinterpretations. We refer to *incorrect* trials twice and for the second time we now have clarified the point *"there should be some preparatory activation detectable in the brain for a starting an incorrect response during incongruent trials but not for congruent trials"*. It is the preparation for responding to the incorrect answer. Any "successful" incorrect response (i.e., they answered, but wrongly) has been taken out of the data and are not part of the analysis, we thus only look at successful trials.

Reviewer 3:

Major:

I have two major concerns that are 1. on the overall rationale and 2. on the analysis steps:

1. your criticise the previous study for the small sample size / power. your study has indeed more than doubled the sample size to 28 (vs 12). Still, I wonder how just argue for this particular sample size as sufficient. It is indeed an improvement, but still as you admit later as well it may benefit from larger sample sizes. Given it is a replication effort, I wonder why you did not do a power calculation to estimate the actual sample size? Given your findings, is it likely that the previous finding with the much smaller sample size is a false-positive?

Reply: The primary aim of the study was not to perform a replication, it was to compare Graves' disease patients with healthy controls as part of a larger study (Holmberg et al 2019). The aim was to investigate activity differences in DLPFC during cognitive fatigability in the patient group during hyperthyroidism, as well as at stable euthyroidism (15 months later). We did however design the study after Szücs et al 2012, and therefore we also measured over the motor cortex. And since we use the same images we though it worthwhile to more in depth compare our results to those by Szücs.

2. Concerning your results: If I understood correctly, you perform a 2 x 2 ANOVA for every single time point. If 10 of those in a row are significant, you calculate the average of those 10 time points and compared the averages again. first, does this then mean you did 120 ANOVAs? If so, how did you deal with multiple comparisons corrections? That seems an extraordinary amount of tests with a high chance of false-positives?

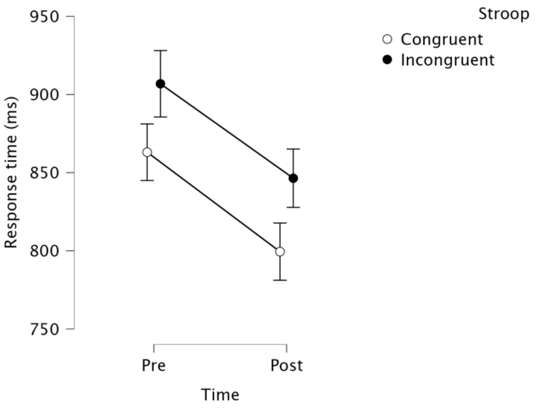
Reply: Yes, this is an accurate description of how we analyzed the data. No additional control for multiple compares is done i.e., change the p-values, but as you state, a change of alpha was done, i.e., we only considered 10 significant timepoints in a row, as a significant time-period. The first question is whether there is a cluster, and when a cluster has been identified it is followed by a second analysis of whether there is a difference in that cluster and what type of difference.

3. you mention you have found a period of 9 time points between 1.8-2.6 seconds for the congruency, that would be below your self-chosen limit of 10 time points. you refer to the initial study that introduced this limit. However, the initial study motivated this in a somewhat arbitrary way: they chose 10 as a higher limit than the potential false-positive rate of 6 readings being significant. I wonder if you could refer to other analysis practices in this regard. I would see this as a somewhat arbitrary limit that you should potentially neglect and use that period for analysis.

Reply: We find this as somewhat contradictory to you last point. We took the spirit of the last point to be that the 10 time points in a row is not sufficient to reduce the risk of false-positive. In this comment we interpret you as saying that our *10-points-in-a-row* rule could have generated a false-negative. We have no calculation to show the relation between false positive rate and false negative rate with this 10 timepoint alpha. So, as such, it is rather arbitrary as you point out. However, the point was to do a replication analysis in this paper, and by the limited space, discussing a change in analysis from the original study would have taken up too much space, and since the main statistic of interest i.e., the interaction, had not a single significant timepoint, we believe it is not relevant for this paper. We do realize that our argument here is based on seeing the result and your concern is about the analysis plan. In the future, we believe it would be better to investigate the peak activity since it would handle the multiple comparison problem as well.

4. I do not see any of the behavioral results that should support your principle assumption of the congruency effect. maybe there would be a chance to correlate the behavioral results to the change in oxy-hb.

Reply: Yes, we agreed with the reviewer. However, due to the space limit, we were not able to add any behavioral result to the manuscript. We have attached some of the behavioral data below. There appears to be a congruency effect, both in the pre and posttest. Response times are also lower in the posttest, but this did not affect the congruency effect.



Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2_p
Stroop	49409.237	1	49409.237	39.617	< .001	0.633
Residuals	28685.161	23	1247.181			
Time	92300.798	1	92300.798	21.026	< .001	0.478
Residuals	100966.302	23	4389.839			
Stroop ✻ Time	61.306	1	61.306	0.120	0.732	0.005
Residuals	11769.238	23	511.706			

Note. Type III Sum of Squares

Minor:

- please review your references, e.g., sometimes the name is D. Szucs vs Denes Szucs. Sometimes there is a full stop after et al. missing

- correct the sentence in the results section: "there here was no significant...". remove a word?

Reply: Thank you for pointing this out, we have now fixed these comments.

Markov Games for Humans and Machines

Claes Strannegård[1,2], **Mattias Rost**[1], Niklas Engsner[2], Johan Lundin Kleberg[3], Mona Guath[4], and Ann Nordgren[2,5]

[1] Department of Applied Information Technology, University of Gothenburg
[2] Department of Molecular Medicine and Surgery, Karolinska Institutet
[3] Department of Psychology, Stockholm University
[4] Department of Psychology, Uppsala University
[5] Department of Biomedicine, University of Gothenburg
mattias.rost@ait.gu.se

Probabilistic games such as the Iowa Gambling Task (IGT) [1] and the Balloon game [2] have been widely used in psychology and neuroscience for clinical research and assessment. We introduce *Markov games*, an infinite family of games containing the IGT and the Balloon game as special cases. The purpose is to provide a uniform framework for testing, training, and modeling probabilistic reasoning in individuals and diagnostic groups. Markov games can be played on digital platforms and a key benefit is that the level of difficulty can be adapted dynamically to the player's performance. Second, we present the results of an initial study with 15 adults playing a selection of Markov games. Third, we consider several computational models based on reinforcement learning and compare their performance to that of the 15 adults.

Markov games

A *Markov Decision Process* (MDP) consists of a set of states S, a set of actions A, a probability distribution $T_a(s, s')$ specifying the probability of moving from state s to state s' when taking action a, and a reward probability distribution $R_a(s, s', r)$, specifying the probability of obtaining reward r when taking action a in state s and moving to state s'. MDPs have been studied extensively in the reinforcement learning (RL) literature [3].

A *Multi-Armed Bandit* (MAB) is an MDP with a single state and a finite set of actions [3]. There are many algorithms with different properties that aim to maximize the accumulated reward of MABs [4]. Such algorithms have been used for a wide range of applications, including recommender systems, network routing, hyperparameter tuning, dynamic pricing, clinical trials, anomaly detection, and influence maximization [5]. MABs have also been used for modeling animal behavior in experiments where animals are presented with different choices and receive feedback in the form of reward, e.g., in the study of bees [6] and pigeons [7]. For example, foraging behavior, where animals choose between exploiting food resources in one area and exploring new areas, has been studied in the case of fruit fly larvae [8].

We define a *Markov game* as a triple (M, s, n), where M is a cycle-free MDP, s is a start state, and n is a positive integer, specifying the number of rounds of the game. To play a Markov game (M, s, n), the player starts in the start state s, chooses actions and receives rewards according to the probability distribution of M (which is not shown to the player), until a leaf of the state graph has been reached. This is repeated n times. The *score* on the game is the sum of the rewards obtained over the n rounds. The challenge of the player is to maximize the score. Examples of Markov games include the IGT [1] and the Balloon game [2]. Four more examples are given in Fig. 1.

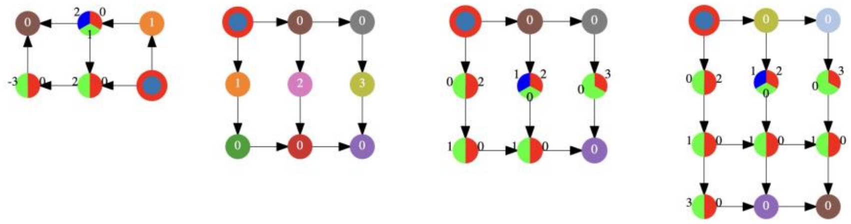


Fig. 1: The four Markov games that were used in the test. The rewards and prob- abilities shown here are not visible to the players. The start nodes are marked with a red circle. The number of rounds was 75 in all four games.

Markov games in different guises have been used for studying humans of different ages and intellectual levels, non-human animal behavior, and RL algorithms. Markov games are easy to construct and vary, with difficulty levels ranging from very easy to impossibly hard. Markov games put cognitive processes such as attention, memory, exploration, and reasoning to test. Understanding randomness, expected value, and calculated risk is helpful for playing the games. A central challenge is to balance exploring and exploiting.

Bandit algorithms

Since Markov games are cycle-free, they might be regarded as MABs, where each path from the start state to a leaf represents an arm. Therefore, we can apply MAB algorithms to Markov games. We used the following classical algorithms:

Random A random algorithm where all paths from the start node to a leaf node are chosen with equal probability.

Greedy The Greedy (Epsilon-greedy) algorithm [3] aims to strike a balance between exploiting the best arm discovered thus far, which it does with a probability of $(1 - \epsilon)$, and exploring other arms by selecting a random arm with a probability of ϵ.

First The First algorithm starts off by selecting random arms for a set of rounds, after which it always selects the best arm following the information it has gathered thus far.

Warm-up The Warm-up (Explore then Commit) algorithm is divided into two phases. In the first phase (Explore), the algorithm pulls as many different arms as possible. In the second phase (Commit), the algorithm consistently pulls the best arm. The difference to the First algorithm is that Warm-up always pulls all arms during Explore, whereas First pulls arms randomly.

Softmax All arms are first pulled once. Then at pull n+1, the arm i is selected with probability

$$p_i = \frac{e^{\bar{x}_i}}{\sum_j e^{\bar{x}_j}}$$

UCB1 In the UCB1 algorithm ~~[4], all arms are~~ first pulled once. Then at pull $n+1$, the arm i that maximizes $\bar{x}_i + \sqrt{2\ln n}/n_i$ is chosen, where $\bar{x}_i$ is the average reward for arm i, n_i is the number of times arm i has been pulled, and n is the total number of pulls.

The Markov games used in the test share two key properties: deterministic actions, meaning an action always leads to the same state, and identical reward distributions for all actions leading to the same state.

Results

We conducted an online experiment with 15 adults with an academic education. Each participant played the four Markov games shown in Fig. 1 and the mean scores were computed. We also computed the mean scores of each of the algorithms mentioned above. The results are shown in Fig. 2 and Table 1. The diagram to the right shows that the five RL algorithms outperformed the human group.

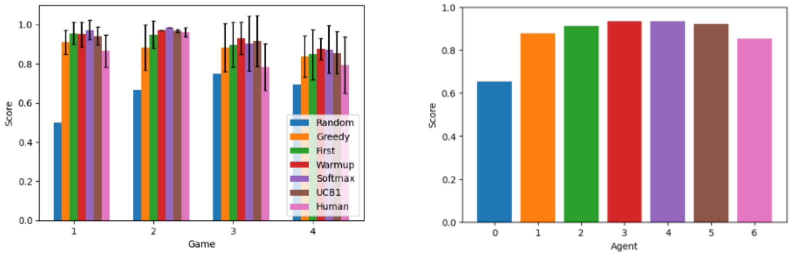


Fig. 2: Performance of humans and algorithms. Each algorithm played every game 100 times and then the mean score was computed. Left: Separated performance on the four games, including standard deviation. Right: Aggregated performance on the four games with the same color coding.

	Game 1	Game 2	Game 3	Game 4
Greedy	4.67E-02*	4.95E-08*	4.62E-03*	2.77E-01*
First	1.06E-04*	1.91E-01	1.18E-03*	2.05E-01
Warmup	2.48E-04*	6.20E-02*	2.30E-05*	3.65E-02*
Softmax	5.00E-06*	1.18E-04*	9.11E-04*	5.42E-02
UCB1	7.90E-04*	2.72E-01	2.25E-04*	1.25E-01

Table 1: Result of two sample T-tests for the scores of RL agents and humans. Positive t-values indicate RL agent supremacy. The symbol * denotes significance (df=113, $p<0.05$).

Conclusion

The results of our initial study suggest that Markov games can be used for testing and training purposes and that player behavior can be modeled using restricted versions of RL algorithms. As a next step, we plan to use Markov games to study various diagnostic groups associated with intellectual disability.

References

1. A. Bechara, A. R. Damasio, H. Damasio, and S. W. Anderson, “Insensitivity to future consequences following damage to human prefrontal cortex,” *Cognition*, vol. 50, no. 1-3, pp. 7–15, 1994.
2. J. L. Kleberg, C. Willfors, H. Björlin Avdic, D. Riby, M. A. Galazka, M. Guath, A. Nordgren, and C. Strannegård, “Social feedback enhances learning in Williams syndrome,” *Scientific Reports*, vol. 13, no. 1, p. 164, 2023.
3. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
4. P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” *Machine learning*, vol. 47, pp. 235–256, 2002.
5. D. Bouneffouf, I. Rish, and C. Aggarwal, “Survey on applications of multi-armed and contextual bandits,” in *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, IEEE, 2020.

6. T. Keasar, E. Rashkovich, D. Cohen, and A. Shmida, “Bees in two-armed bandit situations: foraging choices and possible decision mechanisms,” *Behavioral Ecology*, vol. 13, no. 6, pp. 757–765, 2002.
7. D. Racey, M. E. Young, D. Garlick, J. Ngoc-Minh Pham, and A. P. Blaisdell, “Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules,” *Learning & behavior*, vol. 39, pp. 245–258, 2011.
8. J. Morimoto, “Foraging decisions as multi-armed bandit problems: Applying re- inforcement learning algorithms to foraging data,” *Journal of theoretical biology*, vol. 467, pp. 48–56, 2019.

Reviewer 1. Ronald van den Berg, Department of Psychology, Stockholm University

Contribution is not clear. The contribution of this paper can be stated more clearly. The title suggests that the idea of using Markov Games as an experimental paradigm is the main contribution, but, as the authors indicate on page 2, Markov games have been used before in the study of human behavior. Does the main contribution perhaps lie in the specific variant of the game proposed here? Or in how the data were modelled (Random, Greedy, First, Warmup, Softmax)? Or the combination of the two?

Richness of the paradigm. The authors write that "Markov games put cognitive processes such as attention, memory, exploration, and reasoning to test. Understanding randomness, expected value, and calculated risk is helpful for playing the games.". The richness of the paradigm is rightfully presented as a strength, but I believe that it also presents a challenge.

I happen to have tried the game, after a colleague of one of the authors shared a link with me, because he thought I might find this interesting (neither of us knew at the time that I would be a reviewer on this submission). When I was doing the task, I felt a very strong sense of having to make a trade-off: will I sample more to improve my performance, or will I stop sampling to avoid additional time costs? Perhaps this trade-off is a parameter in the models (and can thus be accounted for in the analysis stage), but if not, the authors might want to look for ways to ensure that subjects make more or less the same trade-off (e.g., by fixing the amount of time per round, so that less exploration does not speed up how fast the experiment is done).

Status of the alternative approach is unclear. The authors write that "Instead of using the average reward for an arm (path) in the algorithms mentioned earlier, an alternative approach is to use the sum of the average rewards for the nodes along the path." What's the status of this proposed alternative approach? Was it tried by the authors? Is it a suggestion for future work?

Confusion about Table 1. The caption mentions t-values, but they are not presented in the table.

Free parameters. Some of the models seem to have free parameters, like the epsilon parameter in the Greedy model. Was the value of this parameter fixed? Or was it adjusted to best describe the data? If so, was this done at the level of individuals, or at the group level?

Model comparison. The authors conducted t-tests to compare human behavior with behavior predicted by each of the proposed models. This approach answers the question "is group-level accuracy distinguishable between the human and between model X?". While this approach has some merit, I think it would be much more informative to do a full, rigorous model comparison (e.g. using ML estimation combined with AIC or cross-validation). That would allow to answer the question "How strongly is model X supported by the data?". Also, since that kind of comparison can be done at the level of individuals, it will also provide insight into individual differences (variation) in cognitive strategies. I'd be happy to have a chat with the authors about this if it's unclear to them how to do this.

Reviewer 2. Felix Thiel, Department of Psychology, University of Umeå

Markov Decision Processes (MDP) are a powerful framework for the modelling of decision strategies in discrete, probabilistic environments. The authors of the present paper are proposing a generalization of the Iowa Gambling Task (IGT) using the multi armed bandit class of MDPs. This is an interesting approach as it allows for comparison between formal algorithmic solutions and human performance in a variety of tasks. Using algorithmic predictions and comparisons to behavior could provide interesting insights into how different clinical groups solve problems.

The authors present the aggregated performance from 15 participants on four tasks which could be modeled as multi-armed bandits. The t-statistics for pairwise comparisons are significant but very small. It is important to keep this in mind when interpreting the results. While this can provide an overview of how human performance compares to specific algorithms, the analysis could have been deeper. The potential of this approach is that participants can be individually fitted to each algorithm, thereby allowing for a better understanding of how one person may have reasoned. Better model fit would suggest that a person may have favored one strategy over others. In turn, this could lead to findings associated with specific clinical groups and a better understanding, treatment, or care of groups of patients.

In light of the potential for fitting individual performance to different strategies, a discussion on the parallels between the described algorithms and human thinking would be a good addition to this paper. Furthermore, an interesting addition to the list of algorithms could be a Bayesian Ideal Observer (IO). The IO optimally integrates information using bayes rule and can provide an upper limit of performance.

On a final note, figure 1 is somewhat unintuitive and could be explained further.

Can a gender ambiguous robot voice reduce gender stereotypes?

Ilaria Torre[1], Erik Lagerstedt[2]

[1]*Interaction Design and Software Engineering, Chalmers University of Technology*

[2]*School of Informatics, University of Skövde*

ilariat@chalmers.se

Introduction

Social robots are meant to be deployed in a variety of places to complete various tasks. When investigating what effect robot characteristics can have on human perception and behaviour, it is reasonable to assume that very often both the embodiment of the robot (i.e., which robotic platform will be used) and its assigned task will be fixed. That is to say, if we want to investigate whether a certain characteristic will influence children's engagement in a learning activity, we will have to work with a fixed robot (whichever robot we have available or have deemed suitable for the investigation) in a fixed context (the learning activity). Both these fixed characteristics can affect how the robot is perceived, e.g. regarding gender. Gender stereotypes have been shown to be transferred to robots (Eyssel & Hegel, 2012). For example, “male” robots were perceived as more suitable for stereotypically male tasks (e.g., repairing technical devices, guarding a house), while “female” robots were perceived as more suitable for stereotypically female tasks (e.g., tasks related to household and care services) (Eyssel & Hegel, 2012; Tay, Jung, & Park, 2014). This poses a dilemma: on the one hand, making sure that robots are matched in terms of gender appearance and task suitability might reduce a potential Uncanny Valley effect (Paetzel, Peters, Nyström, & Castellano, 2016); on the other hand, explicitly mismatching gendered cues might reduce these gender stereotypes over time (Bernotat, Eyssel, & Sachse, 2021). The fact that direct experiences with social robots are still rare makes it particularly relevant to study how human properties are attributed to robots, to minimise deception while remaining understandable (Malle, Fischer, Young, Moon, & Collins, 2020). The initial attributes ascribed to the robots as a kind of stereotypical anthropomorphism can then be incrementally challenged as they are identified and as robots become more common. To be able to balance these phenomena in an informed way, it is necessary to understand how identities are shaped, and how different norms and cues interact with each other. One characteristic that could be explored with regards to gender attribution is robot voice: contrary to robot platform and application context, this feature is “freer”, in the sense that it can be chosen for the robot independently of any hardware or application constraints. Analogous to the previously mentioned dilemma, this raises, on the one hand, the issue of making sure that the voice is appropriate for the robot (see e.g., Moore, 2017). This is sometimes assumed to mean that the voice should be congruent with other physical and social characteristics of the robot, such as its size and shape (Moore, 2017). However, voice may on the other hand perhaps be used to challenge the gender stereotypes afforded by the aforementioned “fixed” robot characteristics. But to reduce these stereotypes we don't necessarily need to add a characteristic that is “opposite” of another one (e.g. a male voice over a female body) and risk enforcing the heterosexuality matrix (Butler, 2002) by exploring identity through its in-congruent cells. Instead we can perhaps try to add a characteristic that is “different”, and try to make the combination of features into something new, while at the same time avoiding dipping into Uncanny Valley territory (Paetzel, Peters, Nyström, & Castellano, 2016). We suggest that one way of doing this would be to use a "gender-neutral" voice. Recently, the Text-To-Speech (TTS) research community has made a few attempts at developing "gender-neutral" artificial voices. One of the initial attempts was Q, a voice generated by a team of Danish activists and researchers in 2019. However, after making some impressions in the media (e.g. https://www.npr.org/2019/03/21/705395100/meet-q-the-gender-neutral-voice-assistant), attempts at contacting the authors of Q to continuing this line of research have failed. Another attempt resulted in the generation of an open-source non-binary voice, generated with several rounds of input from the non-binary community (Danielescu, Horowit-Hendler, Pabst, Stewart, Gallo, & Aylett, 2023). Even more recently, Székely, Gustafson, and Torre (2023), developed synthetic voices by blending recordings of male and female speakers in a neural network, so that the resulting voices were perceived as "gender-ambiguous". A complete review of existing voices is out of the scope of this paper, but these examples highlight the current gap and need for “gender-neutral” or “gender-ambiguous” voices (Sutton, 2020).

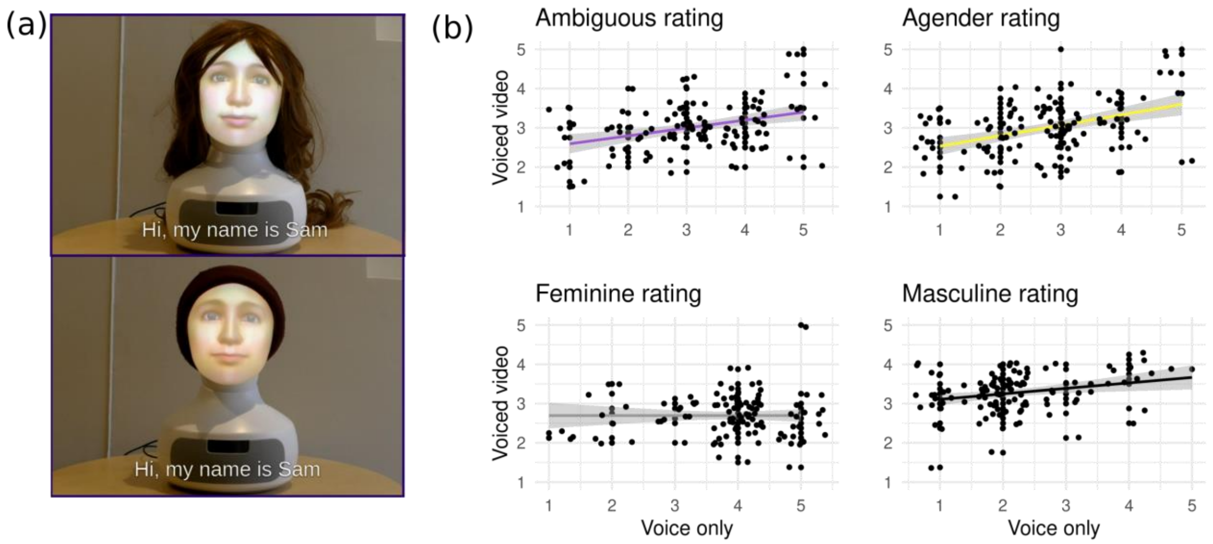


Figure 1. Example of (a), the stimulus (Top; feminine robot, bottom; masculine robot) and (b), linear relationship between gendering of the voice alone and of the voice + robot video.

First results

First of all, we conducted a pilot study online to identify the most appropriate gender-ambiguous TTS voice (chosen from a few current state-of-the-art ones) to use in the main study. This pilot was conducted on Prolific, where we recruited 62 participants (30 identified as female, 31 as male, none as non-binary or preferring to self-describe, and 1 preferred not to say), aged 19--48 (median=25); two participants had to be removed due to failing attention checks. Then, we ran an experiment where participants watched short videos of robots (see Figure 1a) making simple statements in English. 2x2x2 videos were made by manipulating the appearance and “profession” to either be biased to be perceived as more feminine or more masculine, and have each version either voiced with a gender-ambiguous voice or unvoiced (instead texted with subtitles). Each participant watched and assessed each video; half of the participants had the voiced videos first and the unvoiced videos after a distractor task, and the other half got the blocks of stimuli in the other order. For each video, the participant assessed the robot in the video using 4 Likert items (“Feminine”, “Masculine”, “Agender”, and “Ambiguous”), with response options ranging from 1 (= strongly disagree) to 5 (= strongly agree). To ensure common understanding of these terms, we provided participants with working definitions for “Agender” and “Ambiguous”: “By `Ambiguous', we mean that the robot seems neutral or androgynous; by `Agender', we mean that the robot does not seem to have a gender at all”. We recruited 120 participants on Prolific (of which 9 failed the attention check or had technical issues, leaving us with 111 participants). The participants that provided usable data were aged 18--53 years; and 56 identified as male, 50 as female, 3 as non-binary, and 2 preferred not to say. Both the pilot and the main experiment were conducted in accordance with the ethical guidelines of KTH Royal Institute of Technology, and all participants were paid £3.00. The experiment resulted in a rich data set (more details can be found in Torre, Lagerstedt, Dennler, Leite, and Székely, 2023). The results highlight some interesting patterns and phenomena encouraging further research. A specific result worth mentioning here is the relations between the assessments of the voice without any visual cues and the assessments of the voice as it dubbed videos (see Figure 1b). Using linear models to provide inferential statistics, we found significant results in all ratings except for the “Feminine” ratings. More specifically, we fitted mixed-effects linear models and found a significant influence of the voice only ratings for the “Agender” (*b* = 0.27, 95% CI [0.15, 0.38], *t*(109) = 4.72, *p* < .001), “Ambiguous” (*b* = 2.39, 95% CI [2.05, 2.73], *t*(109) = 13.89, *p* < .001), and “Masculine” (*b* = 0.14, 95% CI [0.04, 0.24], *t*(109) = 2.67, *p* = .009) ratings, however, not for the “Feminine” ratings (*b* = 0.00, 95% CI [-0.11, 0.11], *t*(109) = 0.00, *p* = .999). This means that classifying the voice alone as belonging to a specific gender primes people to classify the robot with that voice as having the same gender. This can be seen as partial support for the hypothesis that the gendering of the voice alone will influence the gendering of the robot and voice combination.

Help from the Swedish cognitive science community

With this initial study, we have shown that it is possible to manipulate the gendering of a robot using gender-ambiguous voices, but also that the relations between different factors are complex. Among the future work is to more closely study different subgroups, and to explicitly involve gender-queer people. Additionally, our overarching research question and goal is whether by deploying such a “non-binary” robot long-term in places where stereotypes are formed might result in a reduction of these stereotypes. (Gendered) robots have been present in human societies for a relatively short time. This leaves a relatively open design space, whereby researchers (together with the relevant stakeholders) could explore these ideas and try to break problematic existing conventions via robots. We therefore hope that the discussions at the annual conference of the Swedish Cognitive Science Society will provide further insights towards questions like: (1) Does our approach / research question make sense in a Swedish context? (2) What queer spaces exist in Sweden, and are there any best practices we should employ while involving them in the voice design and evaluation processes? (3) Are there any nuances in gender stereotype formation that are specific to the Swedish context? (4) Have there been any attempts at tackling gender stereotypes in the Swedish context? And (5) what is the role of the institutions—can we get help from schools, local governments, etc.?

Acknowledgments: We wish to thank our collaborators Katie Seaborn, Nathan Dennler, and Éva Székely, who contributed to this ongoing work.

References

Bernotat, J., Eyssel, F., & Sachse, J. (2021). The (fe) male robot: how robot body shape impacts first impressions and trust towards robots. *International Journal of Social Robotics, 13*(3):477–489.

Butler, J. (2002). *Gender trouble*. Routledge.

Danielescu, A., Horowit-Hendler, S. A., Pabst, A., Stewart, K. M., Gallo, E. M., & Aylett, M. P. (2023). Creating inclusive voices for the 21st century: A non-binary text-to-speech for conversational assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Eyssel, F. & Hegel, F. (2012). (s)he’s got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology, 42*(9):2213–2230.

Malle, B. F., Fischer, K., Young, J. E., Moon, A., & Collins, E. (2020). Trust and the discrepancy between expectations and actual capabilities. In Zhang, D. and Wei, B., editors, *Human—Robot Interaction: Control, Analysis, and Design*, chapter 1, pages 1–23. Cambridge Scholars Publishing.

Moore, R. K. (2017). Appropriate voices for artefacts: some key insights. In *1st International workshop on vocal interactivity in-and-between humans, animals and robots*.

Paetzel, M., Peters, C., Nyström, I., & Castellano, G. (2016). Congruency matters-how ambiguous gender cues increase a robot’s uncanniness. In *International conference on social robotics*, pages 402–412. Springer.

Sutton, S. J. (2020). Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity. In *Proceedings of the 2nd conference on conversational user interfaces,* pages 1–8.

Székely, É., Gustafson, J., & Torre, I. (2023). Prosody-controllable gender-ambiguous speech synthesis: a tool for investigating implicit bias in speech perception. In *INTERSPEECH*. ISCA.

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior, 38*:75–84.

Torre, I., Lagerstedt, E., Dennler, N., Seaborn, K., Leite, I., & Székely, É. (2023). Can a gender-ambiguous voice reduce gender stereotypes in human-robot interactions? *In 2023 32th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

Reviewer 1. Lux Miranda, Department of Information Technology, Uppsala University

This is an excellent study doing some very important work. The idea that using robots to reduce stereotypes over time may be of greater value than merely reinforcing those stereotypes to avoid friction in the interaction is well and succinctly argued, and the role of gendered robot voices in this is made very clear! The survey methodology is cathartically adherent to best practices.

I have just a few suggestions to help with the paper:

- I understand that the full details of the statistical analysis are in the RO-MAN paper, but a quick reporting of the p-values for each graph in figure (b) would be helpful in getting a better sense of their significances
- It may be helpful to briefly touch on the difference between ”Agender” and ”Ambiguous;” The difference may be clear to e.g. queer audiences, but I sometimes hear them used as synonyms among the well-intentioned but less gender-savvy
- I’m unsure of the point being made about robots lacking ”historical assumptions regarding potential ’underlying biological structures’”. What are these assumptions? Is the idea that those who have not yet fully accepted trans and nonbinary identity may get hung up on ideas around the coupling of gender and biological sex characteristics, which robots–lacking such biology–may be helpful in breaking down?

In all, this is some great stuff and a well-needed study!

Reviewer 2. Erik Billing, School of Informatics, University of Skövde

This short paper approaches an important challenge in studies of human-robot interaction concerning how specific robot designs affect results, e.g. in the form of Uncanny Valley effects. The aspect considered here is robot gender and specifically the use of gender ambiguous voices in order to reduce stereotypes.

The submitted manuscript unfortunately makes it quite difficult for the reader to understand the details of the study made. The short paper describes both a pilot and a main experiment and after reading the text several times I still have difficulties to untangle exactly how these relate to each other, and from which results are presented. The four correlation plots corresponding to the four items that participants assessed also needs better explanation and clearer axes to be readable.

In sum, I find the present study to be an interesting contribution to the investigation of gendered robots. I also appreciate the questions asked to the SweCog community in the end of the paper. Rather than attempting an answer here in this reflection, I ask the authors to consider how the present work could be taken further, e.g. by investigating how gender assessment by users affect interaction itself?

Towards trustworthy and understandable AI: User-centered Explainability in High-stakes Areas

Shuren Yu[1]

[1]*The Department of Applied IT, University of Gothenburg, Sweden*

[1]*shuren.yu@ait.gu.se*

Abstract

The industry and academia have shown a growing interest in the explainability of AI models. However, a lack of sufficient investigation into the actual needs of users for explainable solutions has resulted in some impractical approaches, leading to situations where even though AI systems provide explanations, users may still not understand the reasons behind the decisions. This may be due to developers ignoring the actual needs of end users while relying on their intuition to provide explanations for AI systems. It is very crucial for users to understand AI decision-making and generate appropriate trust, especially in high-stakess areas. Therefore, this research focuses on conducting semi-structured interviews with AI system users in four high-stakes areas: banking, education, healthcare, and justice. The aim is to understand their needs, satisfaction, and perspectives on the AI explainability in their workflow. This research will provide research direction for AI developers to consider how to build trustworthy and understand AI for users, particularly in high-stakes areas.

1 Introduction

The growth and contribution of AI in various fields are evident, such as banking (Noreen et al., 2023), education (Hu et al., 2018), healthcare (Elemento et al., 2021), and justice (Campbell, 2020). As the practical applications of AI increase, more and more users are concerned about understanding the decision-making and workings of AI to ensure that they showcase their performance in the right way. Understanding how AI reaches conclusions is crucial in sensitive domains, as the consequences of errors can be fatal. As the pursuit of accuracy leads to the increased complexity of models and the technology-centric mindset blindly pursues model performance, understanding the models becomes increasingly challenging. This has sparked discussions on the establishment of methods and strategies for building explainable models.

Regarding the construction of explainability, most current research is based on the principles and techniques of Explainable AI (XAI) (Arrieta et al., 2020). Many studies have developed explainability strategies covering these principles and techniques to provide explanations of AI systems, both in theory and practice (Khan et al., 2022; Dazeley et al., 2021). However, a current issue is that most of these explanations are provided to developers and AI designers, such as heatmaps and analyses of key features, rather than to users in a way and language that they can understand, informing them of the reasons behind AI decision-making. There are two significant reasons for the opacity of AI. First, "writing (and reading) code is a specialist skill" (Burrell, 2016, p.1-2), which most users do not possess. For them, reading these "explanations" is as challenging as reading the AI model itself. Second, there is a "mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation." (p.1-2). The intent and cognition of AI professionals in developing explanations lean towards explaining the logic of machine reasoning in AI decision-making, which does not align with the actual needs for explanations from users, especially those without a technical background, and their ability to interpret these explanations. The underlying reason for this problem may be the cognitive gap between developers of AI explainability and users' understanding of the need for explanations. The evaluation of XAI solutions is usually conducted by their developers, i.e., AI or ML experts, and their 'intuition' about what is a good explanation, rather than with end-users (van der Waa et al., 2021). Designing artificial intelligence that is more in line with users' trust and understanding requires empirical investigations on the specific needs of users for explanations, especially in sensitive domains and high-stakes areas. While there have been several surveys on AI practitioners' needs (Ren et al., 2016; Zhang et al., 2018; Hong et al., 2020), there is still a lack of empirical research on users' perspectives (Anjomshoae et al., 2019; Chazette et al., 2019), particularly understanding their needs, satisfaction, and perspectives on AI explainability in their workflow.

In this paper, I will conduct an empirical study involving twelve AI users in four high-stakes areas (banking, education, healthcare, and Justice) through semi-structured interviews with them. They are professionals in their working field but do not have a technical background in AI. The purpose of my interviews with them is to determine their needs, satisfaction, and perspectives regarding AI explainability at the three levels: conceptualization, construction, and deployment. Subsequently, I will analyze and summarize the data from interviews by a coding approach. I will conceptualize these findings to discuss the research question: *How should AI developers design explainability for users in high-stakes areas to achieve trustworthy and understandable AI?* Finally, I will discuss the limitations, challenges, and future work, that navigate the way for further research on user-centered AI.

2 Related works

In this section, I will first introduce developing explainability at conceptualization, construction, and deployment levels; Secondly, I will discuss existing explainability strategies; Third, I will argue the limitations of explainability methods in practical applications and future work. Finally, I will present the relevant research that motivated me to do this work.

3 Method

My research objectives are as follows: (1) Describe how users understand AI decisions and explainability in four high-risk areas at three levels: conceptualization, construction, and deployment; (2) Summarize and conceptualize users' needs, satisfaction, and perspectives, and answer the research question. To achieve these two objectives, firstly, I will recruit users from these four fields and conduct semi-structured interviews with them; Secondly, I will encode and analyze the interview data; Thirdly, I will conceptualize and summarize these findings. The recruitment process involves convenience and snowball sampling (Creswell & Poth, 2016). The users are all experienced professionals in these four high-stakess areas. Some users were recommended by users who are my acquaintances. They come from 9 different companies or institutions in four areas: banking, education, healthcare, and justice. In the context where they use AI, AI has embedded explainability methods. The participants in these areas have rich experience in their work and most of them have basic knowledge of AI but do not have AI expertise. The design of questions in interviews is guided by the principles of human-centered explainable system designing by Mueller et al. (2021). The reason for using these principles is that they are summarized by Mueller et al. through an extensive literature review on explainable AI, which can help me construct a series of questions about explainability at three levels: conceptualization, construction, and deployment. The interview is semi-structured, and I recorded the voice during the interviews for subsequent analysis with the consent of the interviewees. I also will use professional tools to transcribe the audio into text format, and Atlas.ti will encode and analyze all the textual data.

4 Findings

I will summarize and conceptualize the results of the coding analysis.

5 Discussions

In this section, firstly, I will discuss the findings to answer the research question; Secondly, I will elaborate on the limitations on recruiting participants, the questions in the interviews, and the analysis method; Thirdly, I will discuss how my findings can help calibrate trust; Fourthly, I will argue how to build user-centered explainability strategies; Fifthly, I will describe my future work in user-centered AI research, including broader user surveys, personalized settings for different user groups, and validation of explainability strategies in real-world environments. 84

6. Conclusion

In this research, I will discuss users' needs, satisfaction, and perspectives regarding AI explainability in conceptualization, construction, and deployment by conducting semi-structured interviews. I also will argue how to design explainability for users in high-stakes areas to achieve trustworthy and understandable AI. Finally, I will provide some perspectives and suggestions on trust calibration, user-centered explainability strategies, and future work in user-centered AI.

Reference

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019* (pp. 1078-1088). International Foundation for Autonomous Agents and Multiagent Systems.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1), 2053951715622512.

Campbell, R. W. (2020). Artificial intelligence in the courtroom: The delivery of justice in the age of machine learning. *Colo. Tech. LJ*, *18*, 323.

Chazette, L., Karras, O., & Schneider, K. (2019, September). Do end-users want explanations? Analyzing the role of explainability as an emerging aspect of non-functional requirements. In *2019 IEEE 27th international requirements engineering conference (RE)* (pp. 223-233). IEEE.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, *299*, 103525.

Elemento, O., Leslie, C., Lundin, J., & Tourassi, G. (2021). Artificial intelligence in cancer research, diagnosis and therapy. *Nature Reviews Cancer*, *21*(12), 747-752.

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1-26.

Hu, S., Bhattacharya, H., Chattopadhyay, M., Aslam, N., & Shum, H. P. (2018, December). A Dual-Stream Recurrent Neural Network for Student Feedback Prediction using Kinect. In *2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)* (pp. 1-8). IEEE.

Khan, M. S., Nayebpour, M., Li, M. H., El-Amine, H., Koizumi, N., & Olds, J. L. (2022). Explainable AI: A Neurally-Inspired Decision Stack Framework. *Biomimetics*, *7*(3), 127.

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of explanation in human-AI systems. *arXiv preprint arXiv:2102.04972.*

Noreen, U., Shafique, A., Ahmed, Z., & Ashfaq, M. (2023). Banking 4.0: Artificial intelligence (AI) in banking industry & consumer's perspective. *Sustainability*, *15*(4), 3682.

Ren, D., Amershi, S., Lee, B., Suh, J., & Williams, J. D. (2016). Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics*, *23*(1), 61-70.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, 103404.

Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. (2018). Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, *25*(1), 364-373.

Reviewer 1. Tove Helldin, School of Informatics, University of Skövde

The paper outlines an important area of investigation – despite the many promises of XAI, the community has focused mainly on technical solutions for the explanations and not how these are to be used or how they actually fulfill the needs of the users. See for instance the paper by van der Waa (2021) where it is stated that evaluations of XAI solutions are often performed by its developers, i.e. by AI or ML experts, and their "intuition" of what constitutes a good explanation, rather than together with the actual end users.

The author of the paper explains that he/she will perform interviews with AI-system users, working in five different domains, to extract knowledge about their different explanation needs and, based on that, "fill the cognitive gap" that exists between developers and AI-system users, where the main goal is to "derive user-based strategies for XAI". However, how these interviews will be conducted and what the strategies are envisioned to entail is not provided. I have no doubt that the interviews indeed will be interesting, but to extract the sought-after knowledge based on the described approach (or more lack thereof) will not be possible.

To enable a more generalizable result, the approach for the study needs revisions. The risk is that embarking on the path to interview (future) XAI users, using whatever form of interview strategy you like, you will get individual answers that will be difficult to generalize from. To extract knowledge from the study, you will need better preparation work. For example, according to Moheseni et al. (2020) there are 3 main groups of XAI users – AI novices, data experts and AI experts – all of these groups will most likely need different types of explanations from the XAI implementation as their previous knowledge, tasks etc. will differ drastically. But not only that, users' inclination to trust an AI-based tool often also differs, which will most likely also result in different subjective opinions from the interviewees. Moreover, to ask people "what they would like to know from a XAI solution" might not give you the full story. If instead introducing the users to possible kinds of explanations and how they could be implemented in the user's particular domain, they could be given some inspiration (but also bias) towards preferred explanation types (like how, why, why-not, what-if, how-to, and what-else explanations, again check Moheseni et al.). The evaluation of XAI solution is still under-researched, but there are some papers outlining possible evaluation metrics and approaches, so please have a look at such papers as inspiration for your own approach. Based on those, what will your particular focus be? Explanations that enable a better collaboration between user – AI-tool? Better trust-calibration? Better efficiency? User satisfaction? This needs to be outlined.

Then for the main goal of the study, the lessons learned. It is written in the paper that the goal is to derive user-based strategies for XAI, but what is actually meant? For a particular (sub) user group, working with task X, in need of info Y to do make a trustworthy decision, explanation T is needed? Or what will the strategy entail? Please explain, elaborate and see how to extract that knowledge.

Just two examples of good references to read:

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-45.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence, 291, 103404.

Reviewer 2. Patrick Oden, School of Informatics, University of Skövde

The explainability of AI is an important and, as to date, seemingly difficult problem. The literature is math- and model-heavy. The processes are a bit of a black box. The results are often convincing, but not always accurate or precise. The importance of building trust between researchers and engineers creating the models and the general public, business, and industry users of the models is clear. There is also a sense in some of the literature and especially in popular science that even researchers and engineers do not understand exactly what happens in many of the black box processes.

Recent models seem knowledgeable, precise, authoritative, and believable. The paper by Shuren Yu recognises that the models are sometimes (or often) not these things, and that there may be significant consequences when the models fail to meet expectations. The proposed research also recognises that even when the models produce satisfactory output, without an understanding of how the input results in the output, the output may not be trusted.

The proposed research would contribute to understanding what users need in terms of explainability of AI models. The methodology and motivation seem to be well-considered. There may be a small concern that if the fields that are the targets of this research are too different in foundational concepts and results requirements, then the findings of this research may be somewhat more diluted or less cohesive than if focusing on fewer or highly similar fields where these models are deployed. However, as this is unpredictable before the research is conducted, it may merely have to be a point of interest to look for in the findings themselves.

Review 1

> *Thank you for submitting the paper titled "User- based AI Explainability Strategies: Needs and Challenges". The paper outlines an important area of investigation – despite the many promises of XAI, the community has focused mainly on technical solutions for the explanations and not how these are to be used or how they actually fulfill the needs of the users. See for instance the paper by van der Waa (2021) where it is stated that evaluations of XAI solutions are often performed by its developers, i.e. by AI or ML experts, and their "intuition" of what constitutes a good explanation, rather than together with the actual end users.*

Reply: Thank you for your feedback. You've provided me with important references, and I have already cited the relevant content. This is indeed a crucial research area. The pace of technological development is surpassing our imagination. However, I believe that technology must revolve around people, which is the essence of our research on human-centric AI. Currently, there is a wealth of literature on Explainable AI (XAI), interpretability, and explanatory techniques, but there is a lack of their widespread practical use and, more importantly, a lack of investigation into what non- technical users truly need. Building technology solely based on developers' imagination without considering the practical needs of users can have serious consequences, especially in sensitive and high-risk domains. This is what drives my research in this paper.

> *The author of the paper explains that he/she will perform interviews with AI-system users, working in five different domains, to extract knowledge about their different explanation needs and, based on that, "fill the cognitive gap" that exists between developers and AI- system users, where the main goal is to "derive user-based strategies for XAI". However, how these interviews will be*

> *conducted and what the strategies are envisioned to entail is not provided. I have no doubt that the interviews indeed will be interesting, but to extract the sought-after knowledge based on the described approach (or more lack thereof) will not be possible.*

Reply: Due to difficulties arising from data availability, the originally planned 5 domains have been reduced to 4 high-risk domains, and the number of interviewees is 12 (with 9 interviews already completed). The process of developing an interpretable strategy will be complex and challenging. Too much uncertainty exists before data extraction, organization, and coding are completed, making it impractical to design a strategy. Furthermore, merely developing a strategy does not contribute significantly to the research field because it cannot answer questions like ”Why is my research important?” or ”Why is my strategy superior to others?”. Therefore, this study will focus on gathering the needs, satisfaction levels, and viewpoints of AI decision-making from users in these 4 high-risk domains through interviews. Its purpose is to conduct a more extensive user survey for human-centric AI, especially targeting users with years of experience in high-risk domains who are not AI experts. Based on these findings, the research will address questions such as how to advise AI developers to build user-centric and interpretable AI. The data collection method involves semi-structured interviews, and textual data analysis and conceptualization will be achieved through coding.

> *To enable a more generalizable result, the approach for the study needs revisions. The risk is that embarking on the path to interview (future) XAI users, using whatever form of interview strategy you like, you will get individual answers that will be difficult to generalize from. To extract knowledge from the study, you will need better preparation work. For example, according to Moheseni et al. (2020) there are 3 main groups of XAI users – AI novices, data experts and AI experts – all of these groups will most likely need different types of explanations from the XAI implementation as their previous knowledge, tasks etc. will differ drastically. But not only that, users' inclination to trust an AI-based tool often also differs, which will most likely also result in different subjective opinions from the interviewees. Moreover, to ask people "what they would like to know from a XAI solution" might not give you the full story. If instead introducing the users to possible kinds of explanations and how they could be implemented in the user's particular domain, they could be given some inspiration (but also bias) towards preferred explanation types (like how, why, why-not, what-if, how-to, and what- else explanations, again check Moheseni et al.).*

Reply: As described by Moheseni et al. (2020), the users of interest in this study are considered ”novices” in the four high-risk domains. However, these ”novices” have years of work experience or expertise in their respective fields. Interviewing them can provide better insights into the compatibility of AI systems within their workflow, the usefulness of interpretability, and how to provide effective assistance. It is undeniable that biases in data collection and subjective biases in interpreting the data are inevitable. Still, this does not diminish the intriguing aspects and contributions of this article, namely, how these experts integrate interpretability into their work practices when using AI systems. Such user surveys are valuable for the development of Explainable AI (XAI). Additionally, I will provide a detailed discussion of these limitations in the article.

> *The evaluation of XAI solution is still under- researched, but there are some papers outlining possible evaluation metrics and approaches, so please have a look at such papers as inspiration for your own approach. Based on those, what will your particular focus be? Explanations that enable a better collaboration between user – AI-tool? Better trust-calibration? Better efficiency? User satisfaction? This needs to be outlined.*

Reply: Since this study no longer focuses on constructing a strategy, technique, or method, the interpretation of interview data will primarily revolve around how these seasoned users perceive interpretability. Whether they consider interpretability useful in their workflow, what forms of interpretability they find valuable, and how they perceive the assistance of explanations in their actual work can all help AI developers understand and design user-centric AI. While these data surveys may not establish a standardized AI design blueprint or generalize to all domains, their value, like all other research, lies in providing inspiration for future researchers in this field. This is the research contribution it offers.

> *Then for the main goal of the study, the lessons learned. It is written in the paper that the goal is to derive user-based strategies for XAI, but what is actually meant? For a particular (sub) user group, working with task X, in need of info Y to do make a trustworthy decision, explanation T is needed? Or what will the strategy entail? Please explain, elaborate and see how to extract that knowledge.*

Reply: Constructing a strategy based on interviews with 12 individuals is indeed impractical. Therefore, I have shifted the focus of this study towards extracting insights from users regarding their needs, satisfaction, and viewpoints. By extracting, coding, and conceptualizing this knowledge, I can advise AI developers on what they should focus on. This will help answer questions such as what aspects require attention, what should be incorporated into their product design, and how users can participate in the development process, among others.

> *Just two examples of good references to read:*
>
> *Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-45.*
>
> *van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence, 291, 103404.*

Reply: Thank you very much for the recommendations of these two literature. I have carefully read them and cited the corresponding sections.

> *There may be a small concern that if the fields that are the targets of this research are too different in foundational concepts and results requirements, then the findings of this research may be somewhat more diluted or less cohesive than if focusing on fewer or highly similar fields where these models are deployed. However, as this is unpredictable before the research is conducted, it may merely have to be a point of interest to look for in the findings themselves.*

Reply: As mentioned in the title and main statement, I focus on high-risk areas and collect data from users with rich work experience but no AI background. As I mentioned in response to reviewer 1, constructing a strategy based on a limited number of 12 interviews can indeed be unreliable. Therefore, focusing on interpreting users' needs, satisfaction, and viewpoints may be more practical. These findings can help AI developers think about how to build user-centric AI, especially in high-risk domains where accuracy is of utmost importance, and errors can have fatal consequences. Although the four high-risk domains for interviews are entirely different, they share the common requirement of high accuracy in their outcomes, and mistakes in results and judgments can be critical. These experienced users can provide valuable insights to developers on how to design Explainable AI (XAI) based on their experiences using AI in their workflows and evaluating interpretability. These findings exhibit good cohesion and low sparsity and can potentially be generalized, at least within high-risk domains.