

From DDMs to DNNs: Using Process Data and Models of Decision Making to Improve Human–AI Interactions

Mrugsen Nagsen Gopnarayan¹, Jaan Aru², and Sebastian Gluth¹

¹ Department of Psychology, University of Hamburg


² Institute of Computer Science, University of Tartu

Over the past decades, cognitive neuroscientists and behavioral economists have recognized the value of describing the process of decision making in detail and modeling the emergence of decisions over time. For example, the time it takes to decide can reveal more about an agent's true hidden preferences than only the decision itself. Similarly, data that track the ongoing decision process such as eye movements or neural recordings contain critical information that can be exploited, even if no decision is made. Here, we argue that artificial intelligence (AI) research would benefit from a stronger focus on insights about how decisions emerge over time and from incorporating related process data to improve AI predictions in general and human–AI interactions in particular. First, we discuss to what extent current approaches in multiagent AI do or do not incorporate process data and models of decision making. Next, we introduce a highly established computational framework that assumes decisions to emerge from the noisy accumulation of evidence, and we present related empirical work in psychology, neuroscience, and economics. Finally, we provide specific examples of how a more principled inclusion of the evidence accumulation framework into the training and use of AI can help to improve human–AI interactions in the future.

Keywords: human–artificial intelligence interaction, multiagent artificial intelligence, evidence accumulation, process models, social decision making

From screening mammography for breast cancer tumors with expert-level accuracy (McKinney et al., 2020) to providing more accurate and localized weather forecasts (Sobash et al., 2020), the artificial intelligence (AI) revolution is positively impacting our lives. However, in many of these cases where AI systems perform well, the AI system does not engage in social interaction with a human. Such examples are the solution to the protein folding problem (Jumper et al., 2021), handwriting

recognition (Lecun et al., 1998), and the discovery of novel drugs (Zhavoronkov et al., 2019). In these use cases, humans still set goals by giving labels and setting loss or reward functions, and artificial systems are used as tools to do an asocial task. Yet, there are many contexts such as geriatric care, psychotherapy, or personal assistance, where social interactions are crucial for achieving favorable outcomes, and an effective human–AI interaction becomes a prerequisite to address the

This article was published Online First August 22, 2024.
Mrugsen Nagsen Gopnarayan  <https://orcid.org/0000-0002-5213-8306>

Mrugsen Nagsen Gopnarayan and Sebastian Gluth acknowledge the support of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant 948545). Jaan Aru acknowledges the support of the Estonian Research Council (Grant PSG728) and the Estonian Centre of Excellence in Artificial Intelligence (EXAI), funded by the Estonian Ministry of Education and Research.

Mrugsen Nagsen Gopnarayan played a lead role in writing–original draft and an equal role in project administration and writing–review and editing. Jaan Aru played a lead role in conceptualizing and an equal role in writing–review and editing. Sebastian Gluth played a lead role in conceptualizing and an equal role in project administration and writing–review and editing.

Correspondence concerning this article should be addressed to Mrugsen Nagsen Gopnarayan, Department of Psychology, University of Hamburg, Von-Melle-Park 11, 20146 Hamburg, Germany. Email: mrugsen.gopnarayan@uni-hamburg.de or mrugsenn@gmail.com

issue. In such instances, AI systems are expected to simulate human roles, assisting or coordinating with the user. A competent artificial agent should, therefore, be capable of understanding human behavior, accounting for the likes, beliefs, intentions, desires, and emotions of the person they interact with. This unique capacity is termed theory of mind (ToM). In this review, our goal is to explore existing methods that have been used to try to achieve ToM in AI and to propose a novel approach inspired by decision neuroscience. By focusing on this aspect, we aim to bridge the gap between AI's current capabilities in simpler, nonsocial settings and the demands of socially interactive environments.

In the pursuit of equipping AI with ToM capabilities, lessons can be learned from the functioning of the human brain. This strategy has historically fueled numerous AI breakthroughs. To name just two, the inception of artificial neural networks was inspired by McCulloch and Pitts's (1943) model, which mimics biological neurons. Similarly, convolutional neural networks (CNNs) are inspired by the visual processing mechanisms of the human brain, notably how neurons are structured in the visual cortex, tailoring them for image recognition tasks (Lecun et al., 1998). In this article, we argue that to enhance AI's capability to develop ToM, inspiration can be drawn from how our brain implements ToM in the context of social decision making. Crucially, some current theories indicate ToM to be implemented in the brain through simulating the process of decision making of others (Gallese, 1998; Gordon, 1986). Hence, we argue that understanding general decision-making processes and using process-tracing methods to capture those mechanisms might be the key to ToM in AI.

In the following, we begin by examining the current state of multiagent AI systems and their applications in complex scenarios, such as multi-player games and autonomous vehicle traffic management. Despite their sophistication, we identify the limitations of these systems in achieving true ToM capability. We then shift to discuss decision-making mechanisms in our brain and argue why process-tracing models are better at capturing the nuances of human decision making. Following this, we present an overview of experimental techniques used to observe decision-making processes. In the subsequent section, we discuss how process models and data from process-tracing methods can be utilized to train deep neural networks (DNNs) for simulating human decision

making. We then highlight specific areas where this approach could be transformative and also where this approach may not work. This is followed by a discussion on trust, ethical implications, and potential challenges in implementing this approach.

Progress and Pitfalls in Interactive Artificial Intelligence

When it comes to emulating human social and cooperative behaviors, the classic single-agent machine learning approaches are limited. Conventional single-agent reinforcement learning (RL) approaches such as Q-Learning or policy gradient, for instance, perform poorly as the environments become dynamic and agents change their behavior with training, leading to a moving goal post (Schwartz, 2014). However, inspired by social and behavioral sciences (Duffy et al., 1998) and powered by deep learning, successful multiagent AI approaches have emerged. These approaches, often termed multiagent deep reinforcement learning, involve the use of interconnected DNNs as agents within a system, which are trained using RL principles. Each of these agents interacts with both the environment and other agents in the system to optimize a set of objectives or rewards. Capable of learning to represent complex functions, predicting future states, or estimating the value of different actions, this approach revolutionized multiagent AI (Gronauer & Diepold, 2022). These agents can master complex multiagent video games with superhuman capacity using just simple visual information (Tampuu et al., 2017) and can be adapted to manage the traffic of autonomous vehicles (Zhao et al., 2021).

This level of advancement in multiagent deep reinforcement learning permits the modeling of complex social and cooperative behaviors, akin to those observed in humans and animals (Lowe et al., 2017). Despite the success of AI models such as ToMnet, a DNN that claims to model the agents it encounters solely from observations of their behavior (Rabinowitz et al., 2018), it is debatable whether these networks have truly attained ToM capabilities. For instance, instead of learning the difference between internal states and true states for other agents, an agent like ToMnet might exploit the combination of positions and distances between elements to navigate. This suggests that rather than developing genuine

ToM, DNNs may have learned shortcuts, solving a simpler problem than achieving ToM (Aru et al., 2023). Indeed, in humans, ToM is not merely task-based. For example, children do not repetitively learn to solve the Sally–Anne task to gain ToM. Therefore, instead of training deep learning agents on specific tasks that might require ToM, a better approach would be to train them in more complex environments (Aru et al., 2023).

One solution to the problem would be to learn from the experts in ToM, that is, humans, to acquire this ability. Humans can help agents to learn the policy or reward function, using imitation learning and inverse RL, respectively. For instance, in the realm of autonomous vehicles, where inverse RL is employed, the vehicle learns from observing human drivers (Plebe et al., 2024). But, instead of learning the difference between internal states and true states for other agents, an agent like ToMnet might exploit the combination of positions and distances between elements to navigate. Nevertheless, learning from human preferences (Christiano et al., 2017) or explanation-based learning could bias the agent toward using features with more generalization power and identifying the world's causal structure. For example, the AI model that powers ChatGPT has been significantly enhanced by human feedback loops during its training phase (Ouyang et al., 2022). This iterative feedback process helped shape the model's responses over time, making it more aligned with human values and preferences. ChatGPT correctly solved a percentage of text-based false belief tasks it was tested against (Kosinski, 2023), but see Ullman (2023). This approach can specifically be used to imitate human decision-making processes. AI can gain insights into human preferences, biases, and reasoning strategies by observing and learning from human decision-making patterns.

In real life, people often need to make decisions for others (like buying a gift for Secret Santa) or with others (such as deciding on a location for a team retreat). We can expand on the idea of learning from humans to train AI to the domain of social decision making. Importantly, humans take each other's mental states, preferences, beliefs, intentions, and emotions into account when making decisions in social settings, that is, they use ToM. One approach to ToM in humans is called "simulation theory," which suggests that humans simulate others to understand them (Gallese, 1998; Gordon, 1986). Thus, a person

A could simulate another person B by replaying (or preplaying) the decision that B has made (or is about to make), using A's own decision-making system. Notably, the discovery of mirror neurons that are active both when we perform an action and when we witness someone else performing the same action has been seen as supporting the simulation account of ToM. (Gordon, 1986; Rizzolatti & Craighero, 2004).

This process tracing is key to our simulation of others. Our core proposal is that implementation of ToM in AI would benefit from learning from process models of decision making, which encapsulate a series of cognitive steps that occur during decision making, reflecting the dynamic interplay of perception, evaluation, and action. In the subsequent sections, we will briefly explain these models and describe different tools that can be used to track the emergence of decisions over time before addressing the question of how this knowledge can foster AI development.

Modeling the Emergence of Decisions Using Process Models

Understanding how choices come about requires process models of decision making that describe how decisions emerge over time (from sensory input to motor output). In this regard, human decision making has been modeled as the accumulation of information over time until a decision threshold is met. The drift–diffusion model (DDM) is a widely used mathematical model that represents the accumulation of evidence over time at a specific rate (the drift rate), with some added noise (diffusion; Ratcliff, 1978; Ratcliff et al., 2016). Although being the dominant model in the field, the DDM is only one representative of the much larger class of evidence accumulation models (EAMs; Busemeyer et al., 2019; Smith & Ratcliff, 2004). Importantly, parameters of EAMs reflect latent psychological processes. In the case of DDM, for example, a preexisting bias for either option is modeled by the starting point parameter, the decision boundary represents the cautiousness of the decision maker, while the drift rate models the rate of evidence accumulation. Nondecision time is also considered and assumed to result from sensorimotor processes that are not related to the decision per se. EAMs offer a very general framework for decision making that can be extended to perceptual tasks (Summerfield & Tsetsos, 2012).

Strikingly, a series of recent studies have shown that laypeople seem to intuitively understand a fundamental prediction of EAMs. Specifically, these studies indicate that the time it takes to make a decision usually reflects the decision's difficulty and, consequently, the difference in preference for the options; fast responses are associated with low difficulty and a high difference in preference (Figure 1). In other words, humans take not only decisions but also decision speed into account when inferring others' hidden preferences and beliefs (Bavard et al., in press; Kononov & Krajbich, 2023). The process by which individuals infer others' preferences can be modeled by inverting the DDM and adopting a Bayesian approach (Gates et al., 2021). However, the neural mechanisms of this ability remain elusive.

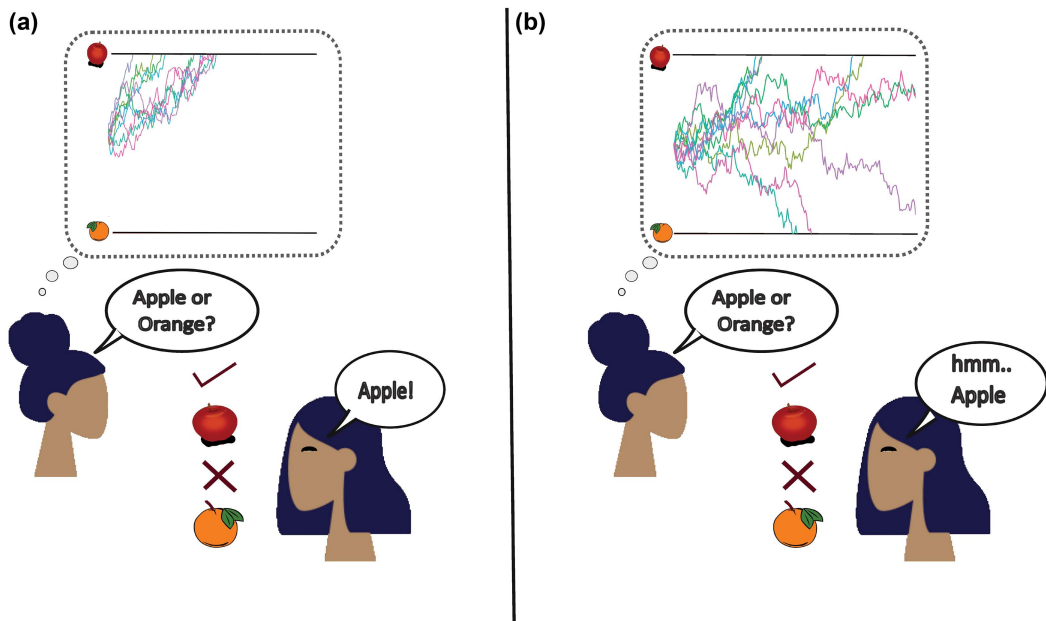
This complex interplay of theory and simulation in human ToM can also be understood as a form of process-tracing imitation. Humans try to simulate scenarios from others' perspectives. This concept provides the motivation for our suggestion that

DNNs should be trained to simulate human decisions. These networks could partially gain the human ability to understand and predict others' decisions, thereby enhancing AI's capabilities in social contexts.

Tracking Physiological and Neural Processes of Decisions

Many advanced AI techniques like neural networks have an immense capacity for learning and pattern recognition, but they also require enormous amounts of data to train. Thankfully, neuroscientists have been using a wide array of techniques to effectively learn and infer the process of decision making. These methods provide a window into the workings of the human brain, offering clues on how decisions emerge and evolve. Hence, using the data obtained from these techniques could be beneficial in training AI systems capable of ToM.

Figure 1
Simulation Theory in Inferring Preferences



Note. The figure demonstrates how humans simulate others' decisions to infer preferences. In (a), a rapid choice for an apple over an orange implies a strong preference for apples. Conversely, in (b), a more deliberative process with multiple simulated scenarios, not all leading to the choice of an apple, suggests a weak preference. The decision trajectories symbolize the cognitive simulation process done by an observer, highlighting how decision latency and pattern can be indicative of preference strength. See the online article for the color version of this figure.

A variety of physiological and neural tools have been used to track the emergence of decisions. Ideally, these measures can be compared with process models of decision making such as the DDM. Although fMRI allows identifying brain regions that are relevant to decision making, the much higher temporal resolution of electroencephalography (EEG) and magnetoencephalography make them more promising candidates to track the dynamic emergence of decisions, including subprocesses such as perception, evaluation, and response preparation (Proudfit, 2014). Contrary to prior belief that these processes happen sequentially (Posner, 1986; Sternberg, 1969), it is now well established that these subprocesses overlap in time and space (i.e., brain regions) to a substantial degree (Gluth et al., 2012; Hare et al., 2011). There is a tight coupling between accumulated evidence, as modeled by EAMs and the activity in the presupplementary motor area. The readiness potential (RP) is an event-related EEG potential that emerges in this region while we are preparing an action (Kornhuber & der Deecke, 1965; Shibasaki & Hallett, 2006). The RP can be understood as a “tendency to respond,” and an experimentally induced sense of urgency leads to an increase in the signal (Gluth et al., 2013). The lateralized RP (LRP) is another neural marker that can be used to investigate the dynamics of action selection. In a task where the choice has to be made using two different hands, LRPs can be calculated as the difference in potentials for left responses versus right responses over the right versus left primary motor cortices. Notably, the onset of the LRP could be an indicator of different nondecision times in decisions with different cognitive demands (memory-based vs. regular decisions; Kraemer & Gluth, 2023). In perceptual decision making, the centroparietal positivity has been proposed as a marker of evidence accumulation (O’Connell et al., 2012), though this claim has recently been challenged (Frömer et al., 2022).

Besides neural mechanisms, there are also powerful tools to track peripheral and physiological signals that are intertwined with decision-making processes. Among these, eye tracking is particularly relevant, since eye movements serve as a window into a decision maker’s attentional processes. Eye movements seem to both influence and reflect preference (Shimojo et al., 2003) and have been used to inform EAMs. The attentional DDM (Krajovich et al., 2010) assumes that fixated

options impact the accumulation process more than nonfixated options. Importantly, many studies have shown that taking eye movements into account via the attentional DDM (or similar models) improves predicting decisions substantially (Gluth et al., 2018, 2020; Krajovich et al., 2010). In addition to eye movements, pupil responses are another critical physiological data source as pupil dilation appears to be a reliable measure of arousal (Joshi et al., 2016) and reveals how decisions evolve (de Gee et al., 2014). Relatedly, choices made contrary to (and thus overcoming) default responses lead to an increased pupil dilation. The starting point in DDM, which represents such response biases, is predictive of pupil dilation (Sheng et al., 2020). Mouse tracking can also reflect the decision process. When individuals encounter conflicting options or experience decision difficulty, their mouse movements can exhibit curvilinear or wavering trajectories, reflecting internal conflict or deliberation (Spivey & Dale, 2006). Heart rate, skin conductance, and cortisol levels are other physiological methods for understanding the decision-making process. However, these measures are more indirect and have high latency; thus, they are not widely used in process-tracing studies.

Although seemingly distinct, social decisions significantly overlap with individual decisions with respect to involved brain regions. Relatedly, it has been argued and shown that the basic framework of conceptualizing value-based choices as emerging from evidence accumulation also applies to social decision making (Gluth & Fontanesi, 2016; Hu et al., 2023; Hunt & Hayden, 2017). EAMs can explain choices and response times in social decision-making tasks as well. Moreover, simultaneous EEG–fMRI recordings have shown that the process of evidence accumulation in the medial prefrontal cortex, which includes the presupplementary motor area, is comparable across social and nonsocial decisions (Arabadzhiyska et al., 2022). However, compared with individual decision making, social decision making is arguably more complex, as it involves understanding and taking others’ preferences or biases into account and learning from others. As a consequence, additional brain regions appear to be critical to social decision making, many of them being said to belong to the ToM network. Central to this process is the temporoparietal junction, a key component

of the theory of mind network. The temporoparietal junction plays a critical role in differentiating one's perspective from others', especially in bargaining situations, and in processing social information and integrating fairness norms during decision making (Hu et al., 2023; Saxe & Kanwisher, 2003). Obviously, humans cannot directly access the neurophysiological information of others when interacting with them. Instead, they need to infer others' thoughts and intentions by observing their behavior, body language, facial expressions, and gaze patterns. Eye movements are especially important as they serve as a window to a decision maker's thought process, revealing their preferences and influencing the outcomes in social interactions. Eye-tracking studies have shown that displaying the gaze allocation of one participant to another in a coordination game improves understanding of each other's preferred choices and facilitates strategic decision-making for maximizing rewards (Hausfeld et al., 2020).

Improving Human–Artificial Intelligence Interactions With Process Models of Decision Making

As discussed above, significant advances have been made to comprehend the computational, physiological, and neural principles of decision-making processes in both individual and social contexts (Kononov & Ruff, 2021). EAMs have been instrumental in this journey, coupled with various tools such as fMRI, EEG/magnetoencephalography, and eye tracking. Recent studies, which focused on how people interpret the decisions of others, have started to shed light on an intuitive human understanding of decision processes that align with EAM principles (Gates et al., 2021). In the following, we build on this work to outline several ideas of how to implement these insights into multiagent AI systems with DNNs. These ideas are thought to help materialize the theoretical understanding of human decision-making processes into tangible AI applications and could help mimic ToM capabilities in AI. We will structure our ideas with specific steps involved in the implementation of process learning in DNN.

Data

Until now, mostly choice data have been used to train neural networks (Kuperwajs et al., 2023;

Peterson et al., 2021). But as discussed in the previous sections, there is a much richer set of data available that contains information about the decision-making process that has not been used thus far. DNNs could also be fed with response times, EEG data, gaze patterns, pupil dilation, or other process-tracing data sets. This will help DNNs to mimic not only the choices made but also the process that resulted in these choices. These data sets may be preprocessed, as is common in most analysis approaches of neural and physiological data in neuroscience and psychology. But preprocessing the data might actually lead to a loss of information and hence should be avoided whenever possible (Khosla et al., 2022). Nevertheless, data sets should be properly formatted, for instance, by converting stimulus into a format that can be processed by a neural network. Additionally, data simulated using process-tracing models could also be used to train the neural network. For example, simulated data using DDMs has been used to train a recurrent neural network (RNN) to match the decision time and the final choice in a probabilistic learning task (Zhang et al., 2020).

Model Architecture

Popular architectures for modeling decisions are RNNs that excel at handling sequential data, making them suitable for modeling decision-making processes that involve a series of steps. Within the family of RNNs, long short-term memory networks and gated recurrent units are particularly effective. They can remember information over longer periods and are less prone to issues like vanishing gradients, which can be a problem with standard RNNs (Shewalkar et al., 2019). Another way of tailoring the architecture is to create a multidomain network, a network comprising different types of layers. For example, a variational layer can be used to capture the inherent variability and uncertainty in human decision making. This layer often employs techniques from variational inference (Oleksienko et al., 2023). It can represent the probability distributions of various latent variables that influence decision making, such as risk preference or uncertainty in outcomes. An attention layer, incorporating attention mechanisms, like those used in transformers, can also be beneficial (Vaswani et al., 2017). These mechanisms help the model focus on the most relevant parts of

the input data, which is similar to how humans focus more on important details. CNNs can also be particularly useful in the preprocessing stage, especially when the input data include spatial or visual elements. CNNs are adept at extracting hierarchical features from images or spatial data. They use convolutional layers to apply filters that detect patterns, edges, and other relevant features.

Tailoring the architecture of the DNN to handle the specifics of process tracing is crucial. One way to achieve this is by mathematically constraining nodes of the neural networks, so that they are functionally equivalent to decision-making models. This has been done where different decision-making models were compared against one another by representing them as neural networks (see Peterson et al., 2021).

Training

The training of the network with process-tracing data is the key step. This might involve a mix of both supervised and unsupervised learning. Supervised learning could be helpful for mapping specific stimuli and physiological data (e.g., sensory stimulus, gaze data, EEG data) to decision-making outcomes (e.g., final decisions, response times), or DNN can be asked to predict the physiological data, ensuring that they would also learn about the process of decision making (Figure 2A). Unsupervised learning, particularly RL, could be applied to tasks that involve learning from environmental interactions to maximize cumulative rewards (like in probabilistic learning tasks).

Validation

The aim of the validation step is to assess how well the neural network replicates human thought processes in decision making. This involves evaluating the model's performance against real human behavior in similar scenarios. One aspect of this will be to test the model's outputs by comparing them with decisions made by humans in identical or very similar decision-making tasks. This could involve using a separate validation data set that the model has not been trained on. Another aspect should be to test AI's performance in novel and distinct choice environments and interactive settings, to test how well the system can generalize to these novel instances. Importantly,

such generalization tests may also allow us to identify whether a multiagent AI system has obtained proper ToM abilities or simply learned shortcuts within a narrow range of choice problems.

Real-World Implications of Process-Tracing Approach in Artificial Intelligence

Given the intricate nature of human–AI interactions, the ability of AI systems to understand, predict, and adapt to human decision-making processes is not just an advantage but a prerequisite for their effective integration into our lives. In the last section, we discussed how process-tracing data and models can be used to train an AI—an approach that could enable AI to mirror and understand the complexities of human decision making. These models, rooted in theory of mind (ToM), will allow AI systems to anticipate human preferences, beliefs, and intentions, with more speed and precision, thereby fostering more natural and effective interactions. In this section, we explore practical scenarios in detail, where we expect the incorporation of process-tracing models into AI to significantly enhance its social intelligence and to improve human–AI interaction.

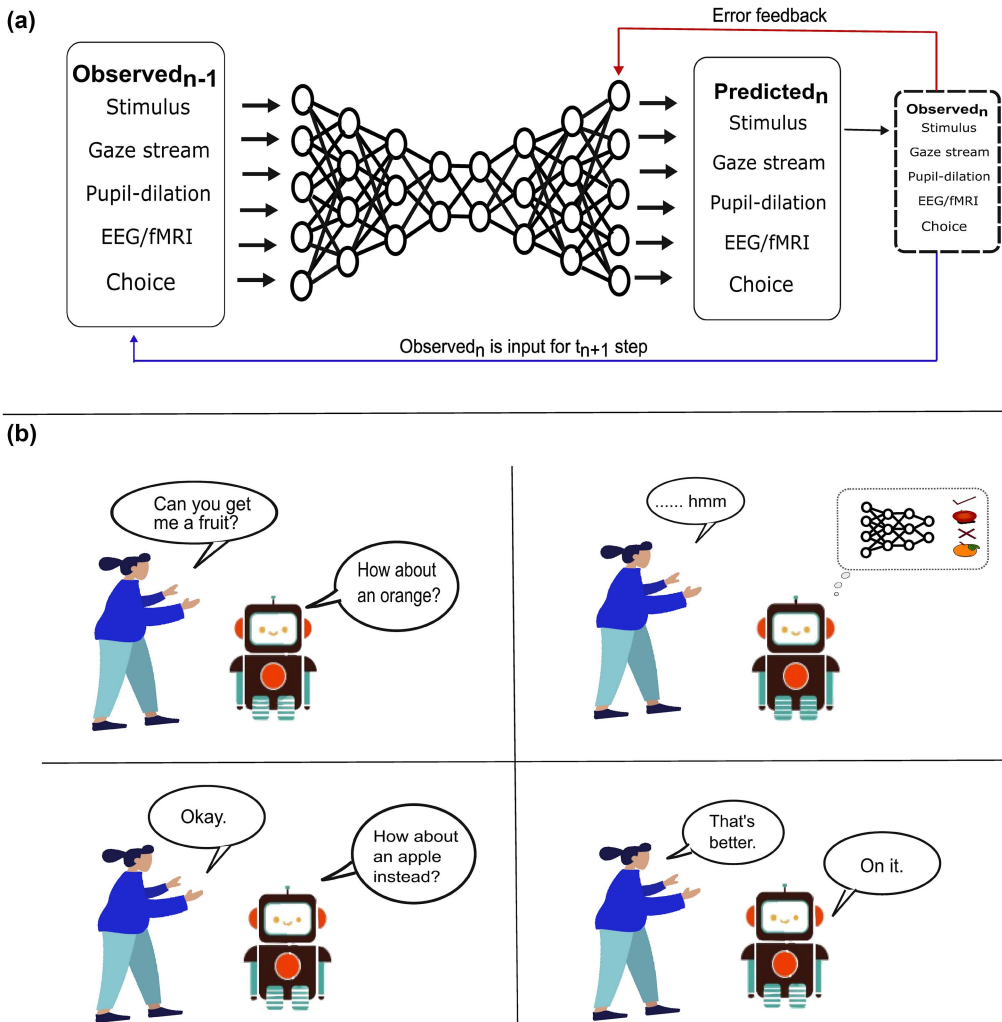
E-Commerce Mouse Tracing

An e-commerce website could utilize mouse tracing to enhance user experience and recommendation accuracy. The website could analyze how users navigate through the site, observing patterns such as the speed of mouse movement, areas where the cursor lingers, and items receiving more clicks. Based on these analyses, the site could infer user interests and indecisions. For instance, if a user frequently hovers over a particular category without making a purchase, the website could deduce a high interest but some level of hesitation, possibly leading to targeted discounts or suggestions of similar but different products to aid in the decision-making process.

House Help Robot

In the future, a robot designed to assist with daily chores, such as cooking or ordering groceries, could use process-tracing models to learn from the past decisions of household members. The robot could also interpret the decision-making process,

Figure 2
Deep Neural Network (DNN) Employing a Process-Tracing Training Mechanism



Note. (a) The figure illustrates a method of training deep neural networks. On the left, observed ($n - 1$) provides the neural network with historical data, including stimulus, gaze stream, pupil dilation, EEG/fMRI, and choice, which are used as inputs for training. The network generates predictions of the same features, which are then compared with the observed (n) data, represented on the right. The comparison yields an error feedback, denoted by the red arrow, which is fed back into the network to adjust and improve the model's predictive capabilities. (b) A person requests a fruit, and the AI robot uses a neural network having a process-tracing mechanism to infer from the person's hesitation that its first suggestion (orange) was not good enough (as indicated by the thought bubble). Using the same network, the AI then identifies a better alternative (apple). This would show AI's ability to interpret and respond to human preferences. EEG = electroencephalography; fMRI = functional magnetic resonance imaging; AI = artificial intelligence. See the online article for the color version of this figure.

where quick choices might suggest a strong preference for certain dishes, and more time spent pondering could indicate a desire for variety (Figure 2B). This learning could enable the robot to make dinner decisions that closely align with the family's preferences.

Supermarket Vending Machines

Supermarkets could be equipped with smart vending machines that recommend products to customers based on process tracing. These machines could analyze customers' eye movements

as they look at different products, interpreting patterns such as the length of time a customer's gaze lingers on certain items or how their eyes move between different product types. These data could help the machines understand customer preferences and indecision, allowing them to make personalized recommendations. For example, if a customer frequently pauses at healthy snacks but usually chooses chocolates, the machine could suggest a sugar-free chocolate bar that combines elements of both.

Besides these encouraging possibilities for enhanced human–AI interactions, there are also scenarios where this approach may be less fruitful. Take, for instance, the field of therapeutics. Human emotions and mental health issues are incredibly complex and varied. Process-tracing models, while sophisticated, might still fall short of fully grasping and appropriately responding to the wide range of human emotional experiences, especially in nuanced therapeutic contexts. Similarly, in the realm of autonomous driving, process models still do not seem to be well poised to capture complicated human behaviors involved in driving and hence may have limited use. However, analyzing the hesitation of drivers before making turns or pedestrians crossing streets could significantly augment autonomous driving systems.

Challenges and Future Directions

Integrating AI systems into our daily lives comes with significant challenges, notably the lack of transparency and interpretability of AI systems. A core problem in this regard is “trustworthy AI,” including questions such as whether and under what circumstances people feel comfortable interacting with AI systems or whether people are willing to transfer responsibilities to machines. Human users often find it difficult to comprehend how an AI system arrives at its decisions, and this lack of transparency leads to reduced trust and acceptance. The major reason contributing to algorithm aversion is the “black-box” nature of these systems (Mahmud et al., 2022). When users do not clearly understand how an algorithm works and how it generates decisions, they tend to exhibit aversion toward these systems, as evidenced by studies (e.g., Dzindolet et al., 2002). People often need justification for algorithmic decisions, and the absence of understandable explanations

reduces their trust in these systems (Lu et al., 2017). Hence, people naturally prefer human decision makers as they can ask them for an explanation behind the decision. The issue of trust is further exacerbated when there is cognitive dissonance due to nonconforming decisions from algorithms (Kitayama et al., 2013).

To elevate this problem, providing clear and understandable explanations of decisions is crucial. When users receive decisions accompanied by explanations of how the algorithm works, they perceive it as more trustworthy (Goodwin et al., 2013). Furthermore, the work of Lankton et al. (2015) draws attention to the intricate relationship between technology, humanness, and trust, highlighting that AI systems imbued with human-like qualities can foster a higher degree of trust among users. Our proposal to incorporate process-tracing models into AI systems can harness this to address the challenge of trust. By mirroring human decision-making processes, these models could offer a more humanlike and interpretable framework for AI systems, potentially improving user trust and acceptance. This could be combined with other explainable AI techniques, such as generating understandable rules or highlighting influential factors. This has shown promise in improving human understanding of AI decisions (Kraus et al., 2020). Effective communication strategies, such as using personalized conversation, illustrations, and persuasive language, can also enhance the perception of trust and acceptance (Yun et al., 2021).

Another challenge in using this framework is the ethical consideration in collecting human data for process training approaches. This approach heavily relies on extensive multimodal data on human behavior. However, the collection and use of such data raise critical ethical questions regarding privacy, consent, and the potential misuse of personal information. Ensuring that data collection adheres to strict ethical standards and respects individual privacy rights is imperative. This involves transparent data collection practices, securing informed consent, and implementing robust data protection measures.

Although this approach of process tracing is quite promising, it necessitates experimental validation to assess its efficacy thoroughly. Initial laboratory studies can be conducted where AI systems are trained using this framework and subsequently evaluated against other approaches. This phase of experimentation is critical to

establish the reliability of the process training approach. Another vital expansion would be the incorporation of models that can capture emotional states. This could be achieved by integrating data from facial expressions, body language, and even physiological signals. By training neural networks to interpret these subtle cues, AI systems can gain a deeper understanding of the emotional context behind human decisions. Applying process-tracing models to more complex scenarios is another critical direction. For instance, in the context of autonomous driving, AI systems can benefit from understanding the decision-making processes of human drivers, such as reaction times, attention allocation, and risk assessment strategies. Finally, continuous improvement and adaptation of process-tracing models is key. As our understanding of human behavior and decision-making evolves, so too should the models we use to train AI systems.

In conclusion, an integration of decision-making research into multiagent AI development holds promising potential for enhancing human–AI interaction. Leveraging process models of human decision making within AI systems can create agents that more accurately reflect and predict human behavior, thus fostering better mutual understanding. These advances can enable the creation of AI systems that are not only more effective but also more transparent and trustworthy. As we continue this exploration, our focus should be on ensuring AI systems that are not just intelligent but also relatable and acceptable to their human users, thereby paving the way for a more symbiotic relationship between humans and AI.

References

- Arabadzhiyska, D. H., Garrod, O. G., Fouragnan, E., Luca, E. D., Schyns, P. G., & Philastides, M. G. (2022). A common neural account for social and nonsocial decisions. *The Journal of Neuroscience*, 42(48), 9030–9044. <https://doi.org/10.1523/jneurosci.0375-22.2022>
- Aru, J., Labash, A., Corcoll, O., & Vicente, R. (2023). Mind the gap: Challenges of deep learning approaches to theory of mind. *Artificial Intelligence Review*, 56(9), 9141–9156. <https://doi.org/10.1007/s10462-023-10401-x>
- Bavard, S., Stuchlý, E., Kononov, A., & Gluth, S. (in press). Humans can infer social preferences from response times alone. *PLOS Biology*. <https://doi.org/10.31234/osf.io/38yrw>
- Bussemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences*, 23(3), 251–263. <https://doi.org/10.1016/j.tics.2018.12.003>
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4299–4307). Curran Associates.
- de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences*, 111(5), E618–E625. <https://doi.org/10.1073/pnas.1317557111>
- Duffy, J., Epstein, J. M., & Axtell, R. (1998). Growing artificial societies: Social science from the bottom up. *Southern Economic Journal*, 64(3), 791–794. <https://doi.org/10.2307/1060800>
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(1), 79–94. <https://doi.org/10.1518/0018720024494856>
- Frömer, R., Nassar, M., Ehinger, B., & Shenhav, A. (2022). Common neural choice signals emerge artifactually amidst multiple distinct value signals. bioRxiv. <https://doi.org/10.1101/2022.08.02.502393>
- Gallese, V. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12), 493–501. [https://doi.org/10.1016/s1364-6613\(98\)01262-5](https://doi.org/10.1016/s1364-6613(98)01262-5)
- Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. L. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition*, 217, Article 104885. <https://doi.org/10.1016/j.cognition.2021.104885>
- Gluth, S., & Fontanesi, L. (2016). Wiring the altruistic brain. *Science*, 351(6277), 1028–1029. <https://doi.org/10.1126/science.aaf4688>
- Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, 4(6), 634–645. <https://doi.org/10.1038/s41562-020-0822-0>
- Gluth, S., Rieskamp, J., & Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, 32(31), 10686–10698. <https://doi.org/10.1523/jneurosci.0727-12.2012>
- Gluth, S., Rieskamp, J., & Büchel, C. (2013). Deciding not to decide: Computational and neural evidence

- for hidden behavior in sequential choice. *PLOS Computational Biology*, 9(10), Article e1003309. <https://doi.org/10.1371/journal.pcbi.1003309>
- Gluth, S., Spektor, M. S., & Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *eLife*, 7, Article e39659. <https://doi.org/10.7554/elife.39659>
- Goodwin, P., Sinan Gönül, M., & Önkal, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, 29(2), 354–366. <https://doi.org/10.1016/j.ijforeca.2012.08.001>
- Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language*, 1(2), 158–71. <https://doi.org/10.1111/j.1468-0017.1986.tb00324.x>
- Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55(6), 895–943. <https://doi.org/10.1007/s10462-021-09996-w>
- Hare, T. A., Malmaud, J., & Rangel, A. (2011). Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. *Journal of Neuroscience*, 31(30), 11077–11087. <https://doi.org/10.1523/jneurosci.6383-10.2011>
- Hausfeld, J., von Hesler, K., & Goldlücke, S. (2020). Strategic gaze: An interactive eyetracking study. *Experimental Economics*, 24(1), 177–205. <https://doi.org/10.1007/s10683-020-09655-x>
- Hu, J., Konovalov, A., & Ruff, C. C. (2023). A unified neural account of contextual and individual differences in altruism. *eLife*, 12, Article e80667. <https://doi.org/10.7554/elife.80667>
- Hunt, L. T., & Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. *Nature Reviews Neuroscience*, 18(3), 172–182. <https://doi.org/10.1038/nrn.2017.7>
- Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, 89(1), 221–234. <https://doi.org/10.1016/j.neuron.2015.11.028>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Khosla, A., Khandnor, P., & Chand, T. (2022). Automated diagnosis of depression from eeg signals using traditional and deep learning approaches: A comparative analysis. *Biocybernetics and Biomedical Engineering*, 42(1), 108–142. <https://doi.org/10.1016/j.bbe.2021.12.005>
- Kitayama, S., Chua, H. F., Tompson, S., & Han, S. (2013). Neural mechanisms of dissonance: An fmri investigation of choice justification. *NeuroImage*, 69, 206–212. <https://doi.org/10.1016/j.neuroimage.2012.11.034>
- Konovalov, A., & Krajbich, I. (2023). Decision times reveal private information in strategic settings: Evidence from bargaining experiments. *The Economic Journal*, 133(656), 3007–3033. <https://doi.org/10.2139/ssrn.3023640>
- Konovalov, A., & Ruff, C. C. (2021). Enhancing models of social and strategic decision making with process tracing and neural data. *WIREs Cognitive Science*, 13, Article e1559. <https://doi.org/10.1002/wcs.1559>
- Kornhuber, H. H., & der Deecke, L. (1965). Brain potential changes during voluntary and passive movements in humans: Readiness potential and reafferent potentials. *Pflüger's Archiv für die gesamte Physiologie des Menschen und der Tiere*, 284(1), 1–17. <https://doi.org/10.1007/bf00412364>
- Kosinski, M. (2023). *Theory of mind might have spontaneously emerged in large language models*. arXiv preprint arXiv:2302.02083. <https://doi.org/10.48550/arXiv.2302.02083>
- Kraemer, P. M., & Gluth, S. (2023). Episodic memory retrieval affects the onset and dynamics of evidence accumulation during value-based decisions. *Journal of Cognitive Neuroscience*, 35(4), 692–714. https://doi.org/10.1162/jocn_a_01968
- Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10), 1292–1298. <https://doi.org/10.1038/nn.2635>
- Kraus, S., Azaria, A., Fiosina, J., Greve, M., Hazon, N., Kolbe, L., Lembecke, T.-B., Muller, J. P., Schleibaum, S., & Vollrath, M. (2020). Ai for explaining decisions in multiagent environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(9), 13534–13538. <https://doi.org/10.1609/aaai.v34i09.7077>
- Kuperwajs, I., Schütt, H. H., & Ma, W. J. (2023). Using deep neural networks as a guide for modeling human planning. *Scientific Reports*, 13, Article 20269. <https://doi.org/10.1038/s41598-023-46850-1>
- Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10), 880–918. <https://doi.org/10.17705/1jais.00411>
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>
- Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017, December 3–9). *Multi-agent actor-critic for mixed cooperative–competitive environments* [Conference session]. Proceedings of the 2017 Neural Information Processing Systems, Long Beach, CA, United States.

- Lu, J., Liang, Y., & Duan, H. (2017). Justifying decisions: Making choices for others enhances preferences for impoverished options. *Social Psychology*, 48(2), 92–103. <https://doi.org/10.1027/1864-9335/a000302>
- Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175, Article 121390. <https://doi.org/10.1016/j.techfore.2021.121390>
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/bf02478259>
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, 15(12), 1729–1735. <https://doi.org/10.1038/nn.3248>
- Oleksienko, I., Tran, D. T., & Iosifidis, A. (2023). Variational neural networks. *Procedia Computer Science*, 222, 104–113. <https://doi.org/10.1016/j.procs.2023.08.148>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 27730–27744). Curran Associates.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision making. *Science*, 372(6547), 1209–1214. <https://doi.org/10.1126/science.abe2629>
- Plebe, A., Svensson, H., Mahmoud, S., & Da Lio, M. (2024). Human-inspired autonomous driving: A survey. *Cognitive Systems Research*, 83, Article 101169. <https://doi.org/10.1016/j.cogsys.2023.101169>
- Posner, R. (1986, December). On the systematics of describing verbal and nonverbal communication. In H.-G. Bosshardt (Ed.), *Perspektiven auf Sprache. Interdisziplinäre Beiträge zum Gedenken an Hans Hörmann* (pp. 267–313). De Gruyter. <https://doi.org/10.1515/9783110886238-016>
- Proudfit, G. H. (2014). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, 52(4), 449–459. <https://doi.org/10.1111/psyp.12370>
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). *Machine theory of mind* [Conference session]. Proceedings of the 35th international conference on machine learning, Proceedings of Machine Learning Research, Stockholm, Sweden.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108. <https://doi.org/10.1037/0033-295x.85.2.59>
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27(1), 169–192. <https://doi.org/10.1146/annurev.neuro.27.070203.144230>
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporoparietal junction in “theory of mind.” *NeuroImage*, 19(4), 1835–1842. [https://doi.org/10.1016/s1053-8119\(03\)00230-1](https://doi.org/10.1016/s1053-8119(03)00230-1)
- Schwartz, H. (2014). *Multi-agent machine learning: A reinforcement approach*. Wiley. <https://books.google.de/books?id=8efoAwAAQBAJ>
- Sheng, F., Ramakrishnan, A., Seok, D., Zhao, W. J., Thelaus, S., Cen, P., & Platt, M. L. (2020). Decomposing loss aversion from gaze allocation and pupil dilation. *Proceedings of the National Academy of Sciences*, 117(21), 11356–11363. <https://doi.org/10.1073/pnas.1919670117>
- Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, 9(4), 235–245. <https://doi.org/10.2478/jaiscr-2019-0006>
- Shibasaki, H., & Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology*, 117(11), 2341–2356. <https://doi.org/10.1016/j.clinph.2006.04.025>
- Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12), 1317–1322. <https://doi.org/10.1038/nn1150>
- Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, 27(3), 161–168. <https://doi.org/10.1016/j.tins.2004.01.006>
- Sobash, R. A., Romine, G. S., & Schwartz, C. S. (2020). A comparison of neural-network and surrogate-severe probabilistic convective hazard

- guidance derived from a convection allowing model. *Weather and Forecasting*, 35(5), 1981–2000. <https://doi.org/10.1175/waf-d-20-0036.1>
- Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, 15(5), 207–211. <https://doi.org/10.1111/j.1467-8721.2006.00437.x>
- Sternberg, S. (1969). The discovery of processing stages: Extensions of donders' method. *Acta Psychologica*, 30, 276–315. [https://doi.org/10.1016/0001-6918\(69\)90055-9](https://doi.org/10.1016/0001-6918(69)90055-9)
- Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience*, 6, Article 70. <https://doi.org/10.3389/fnins.2012.00070>
- Tampuu, A., Matisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., & Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PLOS ONE*, 12(4), Article e0172395. <https://doi.org/10.1371/journal.pone.0172395>
- Ullman, T. (2023). *Large language models fail on trivial alterations to theory-of-mind tasks*. arXiv. <https://doi.org/10.48550/arXiv.2302.08399>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need* [Conference session]. 31st Conference on Neural Information Processing Systems, Long Beach, CA, United States. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- Yun, J. H., Lee, E.-J., & Kim, D. H. (2021). Behavioral and neural evidence on consumer responses to human doctors and medical artificial intelligence. *Psychology & Marketing*, 38(4), 610–625. <https://doi.org/10.1002/mar.21445>
- Zhang, Z., Cheng, H., & Yang, T. (2020). A recurrent neural network framework for flexible and adaptive decision making based on sequence learning. *PLOS Computational Biology*, 16(11), Article e1008342. <https://doi.org/10.1371/journal.pcbi.1008342>
- Zhao, C., Li, L., Pei, X., Li, Z., Wang, F.-Y., & Wu, X. (2021). A comparative study of state-of-the-art driving strategies for autonomous vehicles. *Accident Analysis & Prevention*, 150, Article 105937. <https://doi.org/10.1016/j.aap.2020.105937>
- Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., ... Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37(9), 1038–1040. <https://doi.org/10.1038/s41587-019-0224-x>

Received August 1, 2023

Revision received January 13, 2024

Accepted March 13, 2024 ■