AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

© 2024 The Author(s)
ISSN: 2325-9965

2024, Vol. 11, No. 4, 599–618
https://doi.org/10.1037/dec0000237

Using Artificial Intelligence to Fit, Compare, Evaluate, and Discover Computational Models of Decision Behavior

Peter D. Kvam[1, 2], Konstantina Sokratous[1], Anderson Fitch[1], and Arend Hintze[3]

[1] Department of Psychology, University of Florida

[2] Department of Psychology, The Ohio State University

[3] Department of Microdata Analysis, Dalarna University

Theories of decision making are implemented in models that predict and explain behavior in terms of latent cognitive processes. But where do these models come from, and how are they instantiated in the brain? In this article, we examine several avenues where artificial intelligence (AI) and machine learning (ML) can benefit decision theory by providing new methods for developing and testing cognitive models. First, machine learning can be used to efficiently estimate the values of latent parameters in cognitive models and assign posterior probabilities to competing models of the same observed data. Second, models of decision behavior can be embedded within artificially intelligent systems to allow them to make inferences about human counterparts (goals, abilities, cognition) in real time, equipping AI with tools to interact socially. Third, AI can be used to understand how evolutionary and learning processes give rise to the cognitive abilities that support decision making. Last, the tools of experimental psychology and decision sciences can be applied to better understand the "black boxes" of neural networks by systematically testing input–output (stimulus–response) relationships. Put together, we suggest that merging ML/AI into decision-modeling—and vice versa—is a promising path toward many long-term benefits for both fields.

Keywords: machine learning, artificial intelligence, computational evolution, cognitive modeling

Decision theory—including ideas related to utility, risk-taking, value-based choice, and beyond—is a core concept involved in the design of many artificially intelligent systems (Russell, 2010). Artificial intelligence (AI) was originally conceptualized as a system that could emulate human behavior based on our understanding of human cognition, yet many leaps in AI have come from computational advances rather than advances in our understanding of humans (Gigerenzer, 2023). Despite this, AI and decision sciences are inextricably connected in that we want AI to make good decisions and to help people make better decisions. Even without deliberately building AI based on human cognition, the

This article was published Online First August 22, 2024.

Peter D. Kvam https://orcid.org/0000-0002-3195-8452

This work was supported by a graduate fellowship to Konstantina Sokratous, a SEED grant, awarded to Peter D. Kvam, from the University of Florida Informatics Institute, and a grant from the National Science Foundation (Grant SES-2237119).

This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License (CC BY-NC-ND 4.0; https://creativecommons.org/licenses/by-nc-nd/4.0). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Peter D. Kvam played a lead role in conceptualization, visualization, project administration, supervision, writing–original draft, and writing–review and editing. Konstantina Sokratous played a supporting role in conceptualization, investigation, writing–original draft, and writing–revision and editing. Anderson Fitch played a supporting role in conceptualization, visualization, methodology, writing–original draft, and writing–revision and editing. Arend Hintze played a lead role in methodology and a supporting role in project administration, supervision, writing–original draft, and writing–revision and editing.

Correspondence concerning this article should be addressed to Peter D. Kvam, Department of Psychology, The Ohio State University, 1835 Neil Avenue, Columbus, OH 43210, United States. Email: kvam.4@osu.edu

fields of machine learning (ML) and decision sciences have potential points of contact where the methods of one field can inform research in the other.

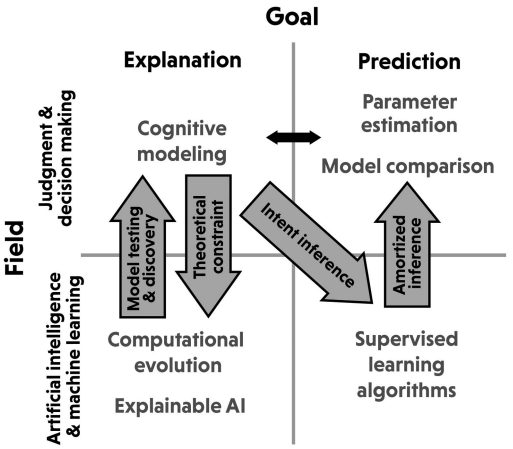
We expect that several contributions to this special issue will examine how AI can be integrated into prescriptive tools that improve the quality of a variety of human choices; in this article, we aim to draw focus instead on how the tools developed in either field can be integrated into the research enterprise of the other. The goal is not to provide an exhaustive account of all areas of overlap between AI/ML and decision sciences but rather to highlight a few particularly promising avenues of research that are productively bringing the fields together. In particular, we focus on methodological innovations in each field that can shed light on problems in the other.

A considerable and potentially fruitful point of overlap between AI/ML and decision sciences is in computational models. However, the goals of modeling frequently differ between fields. In AI, the goal is most often an engineering one: to create a system that can predict an outcome or make a correct classification from a set of input data (supervised learning). Conversely, computational modeling in judgment and decision making (JDM) is more often geared toward explanation—understanding *why* we observe a particular set of behavioral (or neuroimaging/self-report) data in terms of meaningful cognitive processes. They are not limited to these goals, of course, and there are many important cases where models in JDM must carry out prediction (such as model comparison and parameter estimation), as well as cases where models in AI/ML aim to explain a particular process (explainable AI, computational evolution). It is in these crossovers where we suggest there are gains to be made.

An outline of several points of contact between the fields, organized according to the goal of a particular approach or method, is shown in Figure 1. This is not intended to be an exhaustive list but rather an illustration of how we can cross-apply methods according to the goals (prediction, explanation) and fields. The arrows describe what we consider to be promising intersections between the fields and correspond to the four sections of the article, which are organized as follows. The first section examines the practical uses of AI and machine learning as tools for fitting and comparing different theories of decision behavior. Here, AI can speed up the process of parameter estimation and model comparison, opening the field up to new models (theories) that were previously intractable due to a lack of formal likelihood functions. This

Figure 1

Illustration of Two Goals of JDM and AI/ML Research, and Ways in Which They Can Enable One Another Through Cross-Applications



Note. For example, a cognitive model (JDM) can be fit using supervised learning algorithms (AI/ML). Estimates from that model can convey information about latent processes (JDM) to an AI, which may improve its predictions (AI/ML). JDM = judgment and decision making; AI = artificial intelligence; ML = machine learning.

improved prediction ability allows ML-assisted models to be used to rapidly estimate latent processes—such as risk propensity, impulsiveness, loss aversion, or cognitive abilities—from behavioral data. The second section of the article examines intent inference, or how AI can be equipped with cognitive models to better understand human decisions. The ultimate goal of intent inference is for AI to understand the motives and reasons for human behavior, allowing it to better predict and respond to people's actions and desires. The third section then focuses on how AI can be used to evaluate and even discover new models of behavior by simulating the learning and evolutionary processes that give rise to the decision strategies people use. We examine several example cases where AI has been used to test theories of risk aversion, probability weighting, and dynamic decision making by understanding the seed conditions that lead to humanlike behavior. Last, we conclude by examining how models of decision making can be used to understand the relationship between inputs and outputs in opaque models like neural networks, providing theoretical constraints that make it easier to understand and explain what AI and ML algorithms are doing.

Amortized Inference: Model Fitting and Comparison

A cornerstone of judgment and decision-making research is testing and comparing theories by fitting and evaluating different computational models of behavior (Busemeyer & Johnson, 2004; Stevenson et al., 1990; Weber & Johnson, 2009). However, it has become increasingly clear that many effective models do not possess an analytical solution to their probability density (likelihood) function, making it impossible to fit them using traditional methods (Busemeyer et al., 2019; Palestro, Bahg et al., 2018; Turner & Sederberg, 2014). Specifically, prominent methods such as maximum likelihood estimation and Markov Chain Monte Carlo (MCMC) require a known likelihood function in closed form in order to compute gradients and estimate model parameters. For instance, in the example of a linear model, maximum likelihood estimation or MCMC can be applied to estimate the parameters of the regression, such as the slopes and intercept. The Gaussian noise assumption allows us to have a sampling distribution for the estimators because we have an *exact* conditional distribution for each observation (Gaussian) with a known likelihood function. In such cases, models are considered *tractable*. However, for many interesting models of multi-alternative and continuous choice, a modeler must resort to simulation in order to evaluate the frequency with which they predict different patterns of data (Kvam, 2019; Kvam & Busemeyer, 2020; Ratcliff, 2018; Usher & McClelland, 2001), which severely limits their usability in both research and applied settings. To circumvent many practical issues, efforts have been made to develop new methods for complex models, such as likelihood approximation, including probability density approximation, that can be embedded into MCMC/gradient descent algorithms (Holmes, 2015; Turner & Sederberg, 2014).

Approximate Bayesian computation (ABC) is a leading example of simulation-driven inference (Turner & Sederberg, 2012), which seeks to estimate model parameters without requiring traditional likelihood estimation. Precisely, ABC algorithms allow us to sample from posterior distributions over model parameters, which are defined solely as simulators that can be used to create empirical probability densities. An example MCMC process might involve first drawing parameter values from some prior distribution, using these values to generate multiple observations (simulated data sets) from a candidate model, and then rejecting the parameter values for which the distance—or discrepancy—between simulated and observed data exceeds some predefined tolerance/threshold. The process would be repeated until the simulated data were *good enough* in comparison to the observed data. Once that is achieved, the parameter values can be used as approximations of the posterior without any need to calculate the likelihood.

While this example may give the impression that *likelihood-free* algorithms—an umbrella term employed any time an algorithm circumvents explicit likelihoods during estimation—are easy and simple, that is rarely the case (Palestro, Sederberg, et al., 2018). Despite the possibility of using ABC for the implementation of models that were previously inaccessible due to their intractable nature, being able to fit a model does not necessarily mean it is practical or even realistic, given the laborious process one has to go through to achieve that. Most approximation techniques require multiple hours or even days to fit individual participants (see Kvam & Turner, 2021, for some benchmarks with continuous models).

Neural Networks

Fortunately, many issues relating to the efficiency of the aforementioned methods can be sidestepped by using neural networks to estimate the parameters of a model (Boelts et al., 2022; Cranmer et al., 2020; Fengler et al., 2020; Gutmann & Corander, 2016; Lueckmann et al., 2019). Much like ABC and probability density approximation, neural networks can serve as an alternative approximation technique. Like ABC, a neural network-based approach does not require an explicit likelihood, and in many cases, it does not require MCMC sampling either. In this particular case, the algorithm of choice is a neural network trained to carry out model fitting by learning the relationship between observed data and the parameters of a model. The process begins with a step common to all likelihood-free and simulation-based inference approaches, which is simply fixing the parameter space of a model and generating simulated data sets. A modeler creates a training set by first drawing

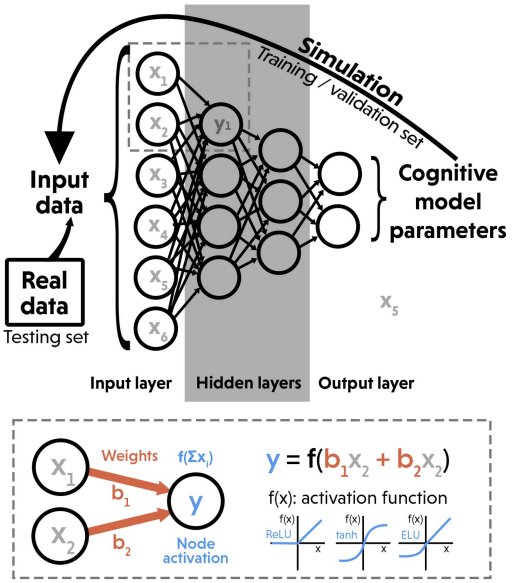
many combinations of parameter values from a particular model (e.g., $m = 100{,}000$ combinations of the n parameters of the model, creating an $m \times n$ output matrix) using a carefully chosen prior distribution, then generating an artificial data set for each combination of parameters by simulating data from the model, and finally using the artificial data and known generative parameter values to train the network on the relationship between the two (Radev, Mertens, Voss, & Köthe, 2020; Rmus et al., 2023).

In an experimental setting, a modeler would only have access to observed data since the true parameters (of complex models) are unknown. Simulation, where the true parameters are known, allows the neural networks to be trained (i.e., simulated binary choice responses from a decision model) to approximate the unknown function relating data (i.e., choices) and parameters (i.e., of the choice model). Essentially, the neural network learns the "task" of mapping data onto parameters. If the parameter space is large/informative enough to produce rich patterns of data sets, then once the function is learned well, the neural network can be used to map observed data onto the parameter values of interest. The reason behind why it can work on any arbitrary model is the universal approximation theorem (Cybenko, 1989; Hornik et al., 1989; Zhou, 2020), a formal proof that stipulates that this process can be accomplished in principle for any model. This means that a feed-forward neural network should be able to approximate the inverse of the simulation function $\Theta \rightarrow D$ (parameters giving rise to data) and map observed data onto parameters, so long as the inverse function $D \rightarrow \Theta$ is approximately continuous and there are a sufficient number of hidden units in the neural network. Once the network is trained on the relationship between data and parameters, it can efficiently estimate the best-fit parameters for a set of input data—such as data from a real participant. A diagram of this process is shown in Figure 2.

In this way, deep neural networks can speed up the process of parameter estimation to a matter of a few seconds or even milliseconds. In our experience, model fitting can proceed at a rate of at least 2,000 participants (inputs) per second for a five-layer, 200-node network—most of which is attributable to data handling used to format inputs to the network. This speed-up arises because the computational burden of the machine learning approach lies primarily in simulating data and

Figure 2

Diagram of How Deep Feed-Forward Neural Networks Can Be Trained and Applied to Estimate the Parameters of Cognitive/Decision Models



Note. Through supervised learning on simulated data with known parameter values, the network learns to estimate model parameters from input data and can be applied to real data. ReLU = rectified linear unit; ELU = exponential linear unit. See the online article for the color version of this figure.

training the neural network. While sizable, the time it takes to do this is orders of magnitude shorter than the time it takes to fit many simulated models using MCMC methods. It, therefore, works best for models that a modeler intends to reuse (usually, particularly common models) and for models that can take a consistent set of inputs, like the quantiles of a response time distribution (Heathcote et al., 2002; Ratcliff et al., 2016). By speeding up the fitting process by several orders of magnitude, it is possible to unlock interesting new practical applications that require model fitting in real time. We review some of these applications in the next section.

Aside from parameter estimation, neural networks can also be used to estimate the posterior variance in parameter predictions (Radev, Mertens, Voss, Ardizzone, & Köthe, 2020), allowing a neural network to estimate not only the best-fit values but also the error in its own predictions. Additional networks can be trained to discriminate

between models—as opposed to merely fitting them—by training a classifier network to take data generated from different models or decision strategies and assign probabilities to generative models (Fang et al., 2023; Radev et al., 2021). Extensions of this approach are beginning to allow for hierarchical model comparison (Elsemüller et al., 2023), fitting models with time-varying parameters (Schumacher et al., 2023), fitting data with missing values using (variational) autoencoders (McCoy et al., 2018), and fitting joint distributions of multiple types of data (Kvam et al., 2024).

With the ability to estimate parameters or generative models in real time, we should be able to embed trained networks in diagnosis systems to better detect medical or cognitive deficits from behavior (Busemeyer & Stout, 2002), give people online feedback about their decisions or performance, and create adaptive training and tutoring programs that adjust their behavior to their understanding of a person's mental states or abilities (Kenny & Pahl, 2009). For example, work by Anderson (1990a) provided an early proof-of-concept into intelligent automated tutoring, using *Adaptive Character of Thought* to construct models of how people execute and acquire skills. By endowing automated tutors with a model of participants' procedural knowledge, this approach could detect when it deviates from the prescribed approach to solving a problem (e.g., geometry proofs or algebra). The breadth of models we can consider in research is also vastly expanded—without the requirement of analytic likelihoods, researchers can consider any model from which they can simulate data. We look forward to the new types of models that are developed using this type of approach.

While likelihood-free methods require simulated data to fit a model, they naturally perform best when the simulated data approximately match the real data. Researchers might favor simulating data from wide distributions over the range of possible parameter values, like uninformative or vague priors in Bayesian estimation. However, neural networks may benefit from training on data simulated from more informative prior distributions. This is especially important when a parameter reaches an extreme range of values that produce the same behavior in the task. Neural networks are robust to mild interpolation and extrapolation, but a high degree of mismatch between a parameter's distributions in the training data and the testing data can make the network's parameter estimates less accurate (Sokratous et al., 2023).

There are also cases where amortized-inference models simply cannot replace traditional likelihood-based models. Proving mathematical relationships between model parameters and data, and the resulting falsifiable empirical predictions, requires analytic solutions to model likelihoods. It is also difficult to benchmark simulated models that have been implemented in deep neural networks, as analytic likelihoods provide a "ground truth" for the maximum likelihood values.

Despite these drawbacks, there are advantages to making simulation-based models more accessible and usable. Many classic models in psychology and decision sciences are how they are because they have convenient mathematical relationships between model parameters and predicted data. For example, the diffusion decision model (Ratcliff, 1978) was inspired by a physical model of heat diffusion. It has its form in part because the solutions to the differential equation describing the diffusion of heat through a material could be applied to predict response times as a function of evidence accumulation. The big advantage of amortized inference is that modelers no longer need to limit themselves to only these convenient types of models. As decision science matures as a field and our understanding of cognition changes, our models will naturally get more complex and less tractable (Schwarz et al., 2009). New theories that are unconstrained by likelihood will allow for new and better explanations, predictions, and questions about the cognitive processes that support decision making (McMullin, 2013). For example, likelihood-free methods can be used to better implement standard models like the leaky competing accumulator model (Miletić et al., 2017), extend existing models with more realistic assumptions like time-varying parameters (Schumacher et al., 2023), or even create and test entirely new theories like dynamic models of pricing behavior (Kvam & Busemeyer, 2020). The value of amortized inference is not only in improved efficiency of estimation and model comparison but also in expanding the scope of models we can consider when theorizing.

Intent Inference: Simulation and Theory of Mind

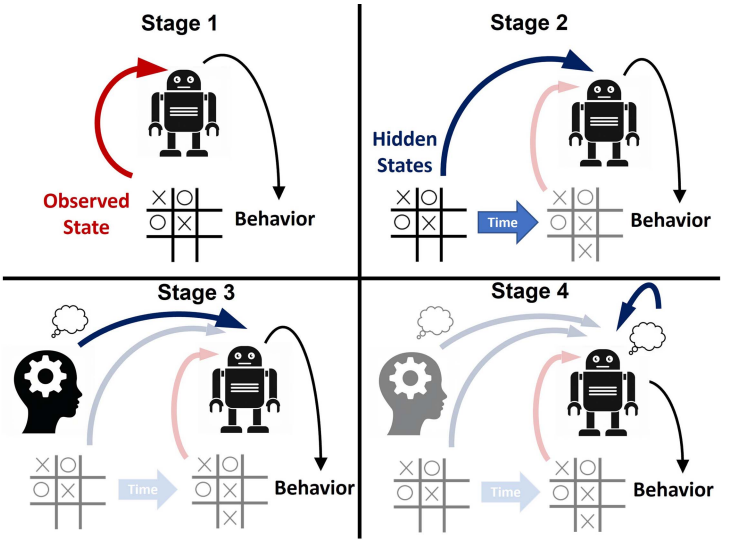
So far, we have discussed feed-forward neural networks, which receive an input, perform a computation, and return an output. These systems

do not retain information about previous inputs and are thus simply mapping an input space to an output space. For many AI tasks such as image classification, this approach is sufficient. However, data from time series often require recurrent systems to accommodate the order in which information occurs and must be processed. Simple recurrent neural networks can accomplish this, but more efficient methods—such as long short-term memory, gated recurrent units, or, to some extent, transformer models—have been developed to perform tasks that require hidden states. To give an example, AlphaZero playing chess does not require knowledge about previous moves, as it does not matter for the optimal move how the current situation on the board came to be. Rather, AlphaZero can make inferences purely by assigning values and probabilities to future states. For other types of AI, such as an AI playing poker, past information might shed greater light on its expectations and assessments of other players; for example, it might be important to know how often or how recently an opponent has bluffed in the past in order to estimate when and how they will bluff in the future.

Interestingly, this distinction between having or not having hidden states might be much more nuanced in natural cognition. Clearly, even simple organisms have hidden states that represent latent motives like fear, hunger, or desired temperature, while they might not have more advanced concepts like curiosity or the urge to produce art. The former constitute primary motives that are directly relevant to survival, while the latter may or may not confer evolutionary fitness—at least not immediately. Clearly, these motives can differ in terms of their complexity and direct relevance to survival or fitness. Thus, a dichotomy between AI systems that have hidden states and those that do not might be too simple. A more refined question might classify cognitive or artificially intelligent systems based on the *content* of the hidden states. A diagram of four potential stages of AI is shown in Figure 3. While these stages of hidden-state representation may not follow this or any particular order, we hope to illustrate a path along which AI systems might develop to become better models for humanlike cognition.

The first stage at which artificial cognition might be situated is machines that does not

Figure 3
Four Stages of AI



Note. In the first stage, an AI responds to the current state of the environment or behavior. The second stage responds to current and past behavior. The third stage responds to behavior over time and the hidden states of others. The fourth stage responds to behavior over time, the hidden states of others, and the hidden states of the self (i.e., metacognition). AI = artificial intelligence. See the online article for the color version of this figure.

possess hidden states and are thus purely reactive. These resemble organisms that only have reflexes or local neural networks, such as the polyp hydra. The second stage corresponds to machines that can retain information about their past environment and thus can form representations on which future decisions can be based. Memory, or at least episodic memory as we recognize it, requires some retention of previously seen information that is not currently available to an organism's sensors, meaning that a system that stores a log or track of its previous states (e.g., keeps track of inputs over time). Some "memory" can be acquired and represented in terms of weights and connections between neurons, but at some point, complex tasks require some sort of information representation above and beyond what is present in their inputs (sensory neurons) and outputs (motor neurons). In AI, systems with hidden or recurrent layers and representations already exist. For example, voice-to-text heavily relies on long short-term memory being able to compare previous sounds with current ones to determine what was actually said and to create a context within which ambiguous sounds can be interpreted. Even insects already have similar capabilities—for example, honey bees remember the spatial locations of flowers based on their flight trajectory and landmarks and can communicate that information to others (Dyer, 2002; Gould et al., 1970; Towne & Gould, 1988).

Theory of Mind

Looking forward, new stages and forms of artificial intelligence can be classified based on what they know about hidden states rather than states of the environment. This is where AI is currently limited and where we suspect it will benefit from JDM's work on cognitive modeling. Specifically, in the next stage of AI or intelligence, more generally, there would be (artificial) organisms that not only retain and infer information about the environment but also about hidden states of other entities. This is also known as the theory of mind, which Premack and Woodruff (1978) defined as having the ability to impute mental states of the self or others. There are some systems that are able to pick up on key parts of this process, such as identifying the visual salience and interactions between different objects or the way in which humans direct attention (Scassellati, 2001), identifying key points for social interaction (Kuniyoshi et al., 2004), metalearning about the behaviors of agents that the AI encounters (Rabinowitz et al., 2018), or attempting to model the mental states of other robots and agents (Sclar et al., 2022). Yet, as we outline below, there is substantial progress to be made.[1]

The last step of developing intelligent machines, and for some researchers, the holy grail of computational neural modeling, would be to have systems that are conscious (Carter et al., 2018; Dehaene et al., 2021; Oliveira, 2022). It is extremely difficult to define consciousness properly, and given the controversies in this domain (e.g., Finkel, 2023; Fleming et al., 2023), we do not seek to do so here. Instead, we can look at some conditions that would make machines more closely approximate the knowledge, capacities, and subjective experiences humans have that are often associated with consciousness. For example, one important property that human cognition exhibits (perhaps as a component of consciousness) is *metacognition*: the ability to form and modify representations of one's own self and thoughts. A being with metacognitive abilities—and arguably, any conscious being—would need to retain information about its own mental and physical state (Descartes, 1901). While this capacity might currently not exist in the form of a computational model or AI, this progression of steps aligning itself to representations about increasingly complex matters suggests a path that AI development might take.

Current AI systems are largely designed to respond to information about past states of an environment to make inferences about future states of the world, placing them around the second stage of machine intelligence described above. For example, *AlphaGo* relies on information about the current and previous configurations of game pieces on a board (past state of the environment) to predict which moves will lead to an advantageous configuration in the future (future state of the environment; Silver et al., 2017). While AlphaGo outperforms even the best human players, it does so without information about who, or what, it is playing against. In contrast, a human player makes decisions using both board configuration and their beliefs about the opposing player and their strategy. This allows a human player to make

[1] Arguably, other systems like poker AI (Yakovenko et al., 2016) also need or benefit from theory of mind when they play against human players, and these systems already exist.

inferences bout why their opponent makes a move—an inference about the hidden states of another's mind—aiding in predicting their next move.

Fundamentally, cognitive models are about making inferences about latent (hidden) mental states from observed behavior (Busemeyer & Johnson, 2004), making them ideally suited to pushing AI into the realm of inferences about hidden states. Such inferences may be about cognition, inferences about a specific person's traits/abilities, and inferences about the underlying algorithms or strategies an individual is using to solve a decision problem (Fang et al., 2023; Liefooghe & Van Maanen, 2023) refers to these as anecdotal, computational, and algorithmic inferences or components of intelligent inference. Theory of mind can, therefore, be incorporated into AI systems through cognitive models (Nguyen & Gonzalez, 2022). An AI system equipped with cognitive models can make predictions about the future states of the world using both environmental information and latent cognitive processes, which, in principle, should increase prediction accuracy compared to environmental information alone.

Potential Applications

To make this more concrete, consider a situation on the road where autonomous and human drivers must cooperate to accomplish their goals. Despite great advances in autonomous driving capabilities, autonomous vehicles struggle to anticipate the behavior of human drivers (Ma et al., 2020). For example, autonomous vehicles must grapple with anticipating and adapting to lane-change behavior. If a human driver begins to move toward an adjacent lane, an autonomous vehicle making inferences only about states of the world might infer simply that the car will be in the middle lane and take no action in response. However, if an exit on the right is coming up, the autonomous vehicle must recognize that a human driver's goal might be to take that exit. In principle, an autonomous vehicle without a theory of mind could know that multilane changes are more likely when approaching an exit, but it would be unable to distinguish which cars are likely to have the goal of reaching the exit. If this same autonomous vehicle was equipped with a cognitive model of behavior, such as approach-avoidance dynamics (Townsend & Busemeyer, 2014), it might be able to recognize that the vehicle is approaching the exit lane and make space in response. As autonomous vehicles become increasingly common, dangerous traffic can be made safer with autonomous vehicles that understand the goals of human drivers and pedestrians that communicate intent through behavior (Matthews et al., 2017).[2]

Beyond fully autonomous processes, the theory of mind may improve AI assistants by incorporating the hidden states of the user and others in the environment (Bello, 2012; Kaber et al., 2005). An AI-assisted prosthetic arm could use information about body posture, opposing hand position, and eye tracking data to make inferences about what action or set of actions a person is trying to pursue with the prosthetic arm (grab, hold, push, throw, etc.; Kidziński et al., 2020; McMullen et al., 2013). More generally, if an AI assistant predicts the goals and strategies of the user and others in the environment (Fang et al., 2023), it can support those strategies or recommend actions that best promote cooperation or compromise. AI poker systems, for example, take the past history of different players and visual cues to their underlying states (e.g., confidence) to make better decisions about gambling and social interactions (Yakovenko et al., 2016). Developing AI assistants for different domains will depend on domain-specific knowledge about human behavior, allowing AI to benefit from research in applied areas of JDM research such as real estate, transportation, auditing, clinical decisions, and business settings (Ashton & Ashton, 1995; Kaplan & Schwartz, 2013; Rohrbaugh et al., 1999).

Of course, identifying when AI has a theory of mind is itself a tricky task. Some have made the argument that large language models can infer human intent, even if they are largely carrying out a process of statistical inference over next-word production (Yildirim & Paul, 2023). However, being trained on such a large volume of human behavior can give it some knowledge of the content that we think underlies that behavior, meaning that its responses can contain nuggets of insight, even if the requisite wisdom does not originate with the large language model (LLM).

[2] To the extent that self-driving cars populate the road, they will also have to understand how to communicate with one another as well. This requires self-driving cars to have hidden states that represent one another's hidden states, and potentially even communicate their goals (hidden states) and their assessments of other nearby drivers to better facilitate road sharing and the harmonious flow of traffic.

This runs into issues like the Chinese room problem (Searle, 1980), where it is difficult to tell or claim whether a computational model "understands" what it is doing. We make a simpler argument, which is that explicitly modeling human cognition will help AI systems better characterize the antecedents of human behavior, providing more accurate, sociable, and coherent AI assistants.

If we are to create intelligent systems with theory of mind, a major goal for decision scientists working at the intersection of AI and cognitive science will have to be modeling the relationships between motivations and actions. Models that predict behavior as a function of plausible individual differences in capacities or goals can be inverted, but only once this relationship is firmly established can AI be equipped with the ability to infer latent states and goals. Identifying the domains that are most important, and where AI can see the greatest gains from intent inference insight systems, will necessarily involve collaborations with industry partners.

Model Testing and Discovery: Computational Evolution

Dating at least back to Simon (1981), decision scientists have speculated about how studying artificial models of the mind can inform our understanding of the human brain and behavior. Using the computer as a testbed for theories, and tinkering with model parameters and structure to identify their predictions and properties, is now a central part of the psychological theory development and testing process (Heathcote et al., 2015). One potential use of AI is to assist in generating and exploring different hypotheses, either by testing many variants of a particular theory (such as prospect theory or delay discounting, Cavagnaro et al., 2016; Peterson et al., 2021), prioritizing search for hypotheses in particularly promising areas (Agrawal et al., 2023), or optimizing experimental designs to identify and explore the most promising possibilities along a continuum of hypotheses (Cavagnaro et al., 2010).

Decision sciences often take a set of constraints or axioms (Savage, 1954) as a point of departure in identifying hypotheses about behavior. Optimal solutions under these constraints, worked out computationally, have been proposed as theories of decision behavior and neural activity (Bogacz, 2007; Meyniel et al., 2015; Tajima et al., 2019; van Ravenzwaaij et al., 2012). However, any optimal decision strategy requires an optimizer in order for a decision-maker to find it in the first place. Typically, it is assumed that (approximate) optimization is carried out by learning and/or evolution (Anderson, 1990b; Drugowitsch et al., 2019; Griffiths et al., 2015; Santos & Rosati, 2015). Here, we focus on the latter, although artificial neural networks can certainly be used to study learning and its interactions with evolution (Hintze et al., 2017, 2019).

Traditionally, the field of AI has focused either on systems that are programed to exhibit certain behaviors or on systems that are programed to learn how to exhibit desired behaviors (Gigerenzer, 2023). However, the progenitor of learning and intelligence, and cognition more broadly, is evolution. Simulating the evolutionary process allows a modeler to explore hypotheses about capacities that cannot be learned or trained in artificial systems. For example, humans' capacities to remember and integrate information over time must have preceded their ability to learn from experience. It is likely that a great number of decision strategies and capacities for supporting decision making were acted upon by natural selection before (or while) they were subject to learning. Evolution can also generate far greater variability in behavior than learning alone—one need only compare the differences in behavior between identical twins (mostly learning) to the differences in behaviors across all the many organisms that inhabit our planet (mostly evolution) to understand the scope of the impact evolution has on behavior. This means that we must understand evolution to understand the complete scope of human behavior. Simulating evolution can, therefore, lead us to more complete theories of behavior and serves as both a research tool for understanding why humans behave how they do as well as a practical tool for creating better artificial intelligence (Back, 1996; Banzhaf et al., 2006; Kvam et al., 2019).

Here, we focus on evolution as a method for discovering and testing decision theories. In particular, we focus on the relationship between evolution and optimality. A key caveat in terms of proposing optimal decision strategies as a basis for models of choice is that global optima are notoriously hard to find, and it is often difficult to even identify what is optimal or adaptive unless a problem is extremely well-specified. Human

cognitive capacities are no exception. Humans have had a long evolutionary history, but many functions like metacognition and higher order reasoning have had a shorter history of selection pressures relative to "older" functions we share with many other organisms. Even capacities that have had a long history of selection have arrived at local maxima because they were previously adapted for physically different environments. For example, human locomotion and vision have been heavily influenced by previous selection for four-legged locomotion and nocturnal conditions (respectively). As a result, even long-standing cognitive and physical structures have inherent idiosyncratic mechanisms, vestigial traits, and "spandrels" that can lower the probability of encountering or achieving optimal behaviors for a particular task globally (Goldsmith, 1990; Krogman, 1951).

Clearly, true optimality is hard to achieve, making it important to study the alternatives. Optimality represents just 1 point in a vast space of behaviors and cognitive mechanisms that use the exactly right information at the exact right time for a well-defined problem; there are many other points in the space of decision behavior that constitute potentially successful or adaptive choice heuristics, particularly for "large world" problems (Brighton & Gigerenzer, 2012). Evolution and learning allow us to explore this space through different mechanisms like feedback, backpropagation, selection, mutation, recombination, and even simple random variation. In this sense, learning and evolution algorithms are helpful on two fronts. First, they allow us to examine what conditions allow learning and evolution to reach an optimal solution and what conditions lead to heuristic decision strategies (Kvam et al., 2019). On the other, simulating evolutionary processes allows us to discover new decision strategies or mechanisms that we might not otherwise have considered as candidate explanations for behavior (Lehman et al., 2018; Ling & Lam, 2019; Miikkulainen, 2021; Miikkulainen et al., 2019). We refer to these approaches as *computational evolution*.

Understanding Cognitive Evolution

This kind of evolutionary optimization results in a sequence of solutions along a trajectory of optimization as agents evolve over generations, often from a random start position towards an optimal or near-optimal solution. Whether and when it actually reaches the optimal solution can be quite variable. In many cases, this dynamic trajectory brings into question the idea of a fixed set of heuristics/adaptive toolbox (Gigerenzer & Todd, 1999), as it suggests that such toolbox might not be filled with a discrete set of separate heuristics, but instead suggests that those heuristics themselves are part of a continuously changing space of solutions that are all subject to evolutionary adaptation and random variation. Instead of taking a strategy or heuristic and examining the evolutionary past that could have led to its generation, computational evolution takes an environment and examines what adaptations agents might evolve in response (Kvam et al., 2019).

Computational evolution further elucidates another question about the optimality or adaptiveness of human cognitive abilities. Specifically, even though human cognitive abilities appear to be among the most sophisticated in the animal kingdom, it is clear that these capacities might not be optimized in an absolute sense—further human evolution could still improve upon what we currently have. As an example, one could argue that probability weighting—specifically, the overweighting of rare events—is a hallmark of an incomplete optimization rather than a "complete" evolutionary adaptation. Similarly, it could result from optimization under constraints that are the product of evolution—such as costly information search, limited memory, or strong priors (Vul et al., 2014).

The potential insights generated by computational evolution are ironically rooted in its weaknesses as an optimization algorithm. Evolution is more than a route to the optimal solution; it has a wide range of idiosyncracies that can have an impact on the development of cognitive abilities. These idiosyncracies are core components of biological systems: the way mutations alter strategies confounds the trajectory of optimization (Adami et al., 2016; Schossau & Hintze, 2020); the computational (biological) substrate used to implement cognitive mechanisms profoundly confounds any evolved solutions (Hintze et al., 2019); and the number of environmental parameters that could guide evolution is innumerable. This complex interdependent nature of cognitive evolution superficially appears as a barrier to psychological inquiry, yet the careful and controlled investigation of conditions that lead to the emergence of different decision strategies is the strength of this approach.

Instead of merely positing that a particular cognitive capacity has evolved, we can pinpoint conditions that were conducive or prohibitive to its evolution. For example, work by Hintze et al. (2015) showed that risk aversion—typically believed to be irrational—actually leads to better long-term fitness over many generations by preventing populations of risk-takers from dying out. However, its benefit is typically realized only in the small population and group sizes humans have experienced, tying the evolutionary conditions under which risk aversion could evolve to the evolutionary conditions of human history.

This type of approach is most effective when testing the (relative) plausibility of particular explanations for human behavior. If we test two evolutionary environments, X_1 and X_2, and behavior Y emerges only in X_1, then we might assign greater credibility to X_1 as an explanation for why Y emerged. We will never definitively conclude what environment gave rise to a particular behavior we observe in humans, but we can at least test the relative veracity of different proposals in a Bayesian way (Kwisthout & Van Rooij, 2013). This allows us to carry out abduction and theory comparisons (Blokpoel et al., 2018), arriving at better and better explanations for human behavior as the approach is iteratively applied.

Consequently, using computational evolution to understand the conditions under which decision making evolved allows us to generate new testable hypotheses, including both psychological hypotheses about what cognitive mechanisms might support decision making as well as biological and anthropological hypotheses about how these mechanisms came to be in the first place. It, therefore, helps in motivating and understanding both the past and present of human decision making.

Discovering New Theories

Unlike aforementioned machine learning techniques that find solutions based on a preexisting level of knowledge (LeCun et al., 2015; i.e., train a network to classify whether a picture depicts an animal or an object), computational evolution does not require a presupposed notion of a correct answer or global optimum. This mimics the properties of our evolutionary past—there has never been a target solution to find, but rather genetically "nearby" neural structures that might give rise to better decision mechanisms (judged in terms of greater "fitness" like higher rates of survival, finding food, procreating) than the current ones. By simulating the evolutionary process with reasonable seed conditions, we can, therefore, identify paths that human evolution might have taken, along with the decision strategies that might have evolved under different combinations of conditions. The closer our evolutionary conditions are to the human evolutionary past, the closer we can expect the resulting strategies to align with real behavior (Hintze et al., 2017).

Another feature of evolution that is not realized in traditional gradient descent or other optimization algorithms is its ability to change the problem under consideration and the dimensionality of the problem. Rather than optimizing a set of parameters under a set of constraints, evolution can change the constraints—for example, by introducing a new type of sensory neuron or a new cognitive capacity—or it can change the dimensionality of the parameter space (e.g., with insertion or deletion mutations). This makes evolution fundamentally different than most optimization algorithms, as it can generate previously unknown innovations to a known problem in an agent's environment, or it can solve a different problem altogether.

As a result, computational evolution offers an alternative approach to optimization for developing new theories of decision making. It entails looser, yet more explicit, assumptions about how decision strategies came to be. Any evolved structure generated in this way need not be optimal and, in fact, is likely not to be. Yet incorporating evolutionary components into our models of behavior brings us closer to the trajectory of development that gave rise to human decision behavior. Its utility lies not so much in identifying the one true way that behavior came to be but in deriving new hypotheses and explanations by virtue of integrating more realistic assumptions into the way that we come up with and test hypotheses. In this way, computational evolution can be seen as a "bottom-up" approach to theory building; rather than taking a problem, solving it, and then looking for cognitive and neural architectures resembling the solution. Instead, we take a problem and allow evolution to generate its own solution to the problem. So long as the evolved substrate resembles realistic structures—consisting of neurons, connections, a base-pair genome, and so on—we are guaranteed a

biologically plausible solution to the task. This enables our approach to hypothesis generation and testing to embody more elements of "abduction proper" (Blokpoel et al., 2018) that contribute to good theory development.

The next step, then, is to identify what elements of human behavior are captured by the evolved solutions to a task. To the extent that they do not, we can modify either the structure of the evolved agents or the environmental constraints we believe humans might face during the problem—following Simon's dual constraints of the structure of the environment and the computational capabilities of decision-makers (Simon, 1990). To the extent that evolved agents reproduce patterns of human decisions, we can use them as novel models for studying and understanding the structure of the mind, brain, and their relation to behavior. In the same way that model organisms are used to study particular components of our biology and behavior—whether we are looking at *E. coli* or apes—artificial brains and artificial organisms can help us learn about the roots of human evolution and its relation to our brains and behaviors.

Theoretical Constraints: Behavioral Experiments on AI Systems

Whether using computational evolution or machine learning to create an AI system, a common hurdle is understanding what exactly the AI does. Many types of neural networks are notorious for being "black boxes"—fully connected networks are particularly difficult to understand because the role of any particular node in the network is characterized by its interactions with all of the nodes in the previous layer and all of the nodes in the successive layer. This makes it difficult to predict when and where errors will occur (Castelvecchi, 2016), as well as explain to nonexperts how and why the network is working. This lack of explainability is particularly important when it comes to human–AI collaboration and interaction. A person's trust in AI depends not only on performance but also on factors like explanations, predictability, and communications of confidence (Yu et al., 2017; Y. Zhang, Liao, & Bellamy, 2020). Efforts at creating explainable AI (Gunning et al., 2019), especially with complex systems like deep learning models, have previously used methods of explanation based on selected examples designed to illustrate why a model makes a few of its predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016). However, these approaches are still unable to elucidate the relationship between network inputs and outputs that many users might prefer.

Cognitive science and experimental decision sciences are uniquely positioned to create methods for understanding black-box systems and creating better, more explainable AI (Cassenti et al., 2022; Taylor & Taylor, 2021). These fields specialize in explaining the relationship between inputs (sensory processes) and outputs (behavior) in terms of a set of coherent latent processes, often in the form of cognitive modeling (Busemeyer & Johnson, 2004). Even in cases where these models are incorrect, or cases where they are unfalsifiable (Jones & Dzhafarov, 2014), models can still provide a common language for describing the behavior of a cognitive system, whether human or artificial (Shiffrin, 2010).

Understanding what a (deep) neural network is doing can, therefore, be treated as a problem of abduction, just as in the case of human cognitive systems (Josephson & Josephson, 1996). The idea is to take the behavior of the system and come up with a parsimonious and useful explanation for what it is doing. This is a case where the modeling tools that are used to study human behavior can be applied to understanding the behavior of the AI: we identify theories of behavior on the task, generate predictions from each theory, titrate the inputs to the neural network in order to produce interesting behavior, then examine which model best accounts for its behavior. For example, the work by Kvam and Hintze (2018) examined the behavior of artificial neural networks by comparing heuristic and (optimal) sequential sampling decision strategies for reward rate optimization problems. There, the authors found that different environmental conditions led to decision strategies that were better described as sequential sampling (Ratcliff et al., 2016) or better described as heuristic strategies like run rules (Fific & Buckmann, 2013). Of course, few agents actually implemented these exact algorithms, yet describing them in terms of known or hypothesized decision processes made it much easier to describe and understand what the evolved neural networks (in this case, Markov brains; Hintze et al., 2017) were doing.

The key takeaway here is that model comparison is useful for understanding behavior, even when

we know the models under consideration are all wrong because it describes that behavior in an interpretable common language. The true model is one characterized by nodes, weights, activation functions, and other transformations, yet we can still understand the aggregate behavior of a neural network in simple terms. This is similar to understanding cognition in terms of different levels, as done by Marr (1982). We know the implementation and (to an extent) algorithmic levels by virtue of understanding the pieces of the neural network—yet description at the computational or representational level is more important for the purposes of explaining its behavior in a comprehensible way (Miller, 2019).

Last, we can try to understand complex systems by attributing individual components to the different functions that facilitate them. In biological systems, this is typically done by ablation or knock-out studies linking a neuron or gene to a specific phenotype. Computational systems, which allow us easy access to all components, allow us to use information theoretic tools to dissect the functions of different parts of autonomous systems by doing the same (Hintze & Adami, 2023).

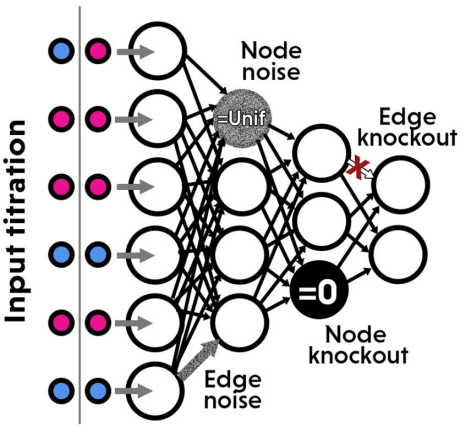
Types of Manipulations

Decision theorists actually gain a number of advantages when moving from studying human systems to studying artificial systems. For example, one can control the entire testing environment, isolating the importance of different input features by titrating them independently and weighing their impact on network outputs (Z. Zhang et al., 2018)—a feat not possible even in highly controlled laboratory studies with human participants. Furthermore, an experimenter can produce behavior in a fraction of the time, allowing for a much greater number of trials and experimental manipulations.

The ethical and practical considerations of manipulating neural activity also disappear when moving to artificial neural networks. Typically, understanding the roles of different parts of the brain and manipulating neural activity require extremely coarse (and often expensive and time-consuming) measurements and manipulations like functional magnetic resonance imaging, positron emission tomography, magnetoencephalography, electroencephalogram, transcranial magnetic stimulation, and transcranial direct-current stimulation (Shafi et al., 2012). Comparatively, interventions on artificial brains can be much more precise, as we can measure and manipulate the activity of every individual node in a network. The accessibility of the decision-making machinery in AI allows the experimenter to extract richer and more precise data. In a neural network, we can record and analyze every hidden state of every neuron, every connection weight, and how they respond to nearly any set of valid inputs. Such data readily permit the identification of correlations between the computational components and the underlying decision-making process (Hopfield & Tank, 1985). However, a wide range of perturbation assays allows for a more experimental approach. We can establish a causal relationship between the output of the decision-making process and individual components by directly knocking out or noising a component of the network. Some of these manipulations are shown in Figure 4. If the component, like the value of a specific weight, is relevant, the output will be broken or as perturbed as the noise is applied to the connection, whereas manipulating a redundant or irrelevant component will have no effect on behavior (Hintze, 2021; Hintze et al., 2022). Studying the effect of perturbing pairs of components allows the unveiling of even more complex, so-called *epistatic* interactions (Bateson & Mendel, 2013).

Figure 4

Diagram of Several Types of Manipulations That Can Be Carried Out on Neural Networks in Order to Experimentally Understand the Functions of Different Nodes, Connections, or Subnetworks/Subsystems



Note. See the online article for the color version of this figure.

For example, a common mechanism in neural networks that instantiates what we might consider "short-term memory" is a chain or recursive sequence of nodes that pass the same information from node to node. If neurons or nodes are labeled with letters, this information might go $A \rightarrow B \rightarrow C$ or even $A \rightarrow B \rightarrow C \rightarrow A$. To determine whether this circuit is involved in storing information relevant to a particular task, we might perform either a knockout or noise analysis on one of the nodes, fixing B to zero or drawing its value randomly rather than allowing it to receive input from A. Alternatively, we could distort the weight of the connection between $A \rightarrow B$. Either approach allows us to examine whether B or its inputs are important to accomplishing a particular task. For example, if B stores and passes along information during evidence accumulation, manipulating it or its inputs would reduce accuracy and potentially even prevent the neural network from making a response (or result in faster, guess responses depending on the input). By manipulating nodes, weights, and inputs, we can experimentally examine the role B plays both descriptively and in a model-based theory of how the network works.

Information-theoretic approaches complement this correlational analysis and have been used to characterize cognitive and computational systems alike (Hintze & Adami, 2023; Marstaller et al., 2013; Tehrani-Saleh & Adami, 2020; Tononi, 2004) and can also be combined with perturbation analysis (Bohm et al., 2022; Hintze & Adami, 2022). Here, instead of relying on correlations, information theory allows us to directly quantify how much more another variable improves the predictability between already established ones. Similarly, the amount of information passing through a system, or how much information needs to be stored in hidden states in order to predict output states, can be quantified (Ay & Polani, 2008). This allows us, for example, estimate the amount of memory needed to complete a task or the amount that a particular piece of information impacted an artificial decision-maker's choices. Using these tools, a decision theorist can carry out experiments, sophisticated manipulations, and data analysis on artificial systems in a way that makes them much more transparent and explainable.

Naturally, there are limitations to what exactly we can know, even with the benefit of causal knowledge about how individual nodes or weights in a hidden layer affect the behavior of a neural network. Often, these types of interventions elucidate the weight of a particular node or connection, but not its function. In some ways, this is a drawback of the distributed representations present in neural networks with fully connected layers: information is "smeared" across many nodes, leaving each individual node of a hidden layer without a coherent role. This is another place where evolved systems differ from those typically used in AI: real brains tend to be more modular and sparsely connected compared to artificial neural networks (Happel & Murre, 1994). Using evolutionary algorithms that specify the presence or absence of connections rather than just their weight (Hintze & Adami, 2022; Hintze et al., 2017), creating and testing modular systems that perform specific functions (Bryson, 2005; Ellefsen et al., 2015), and drawing on the structure of the brain and biological systems (Cox & Dean, 2014; W. Zhang, Gao, et al., 2020) are likely to lead to more effective neural networks as well as artificial systems that more closely resemble the human brain.

Discussion

There is a great deal of work left to be done in each of these areas, and fortunately, an increasing number of decision researchers are taking an interest in machine learning methods. Some have been using machine learning to discover or approximate the shape of functions that are relevant to decision making, such as utility, probability weighting Peterson et al. (2021), and temporal discounting functions (Cavagnaro et al., 2016). Others are using natural language models like vector-based approaches or transformers (Bhatia, 2023; Bhatia & Mullett, 2018) to directly make predictions about human behavior. And yet others are discovering ways to decode neural representations during decision making (Horikawa et al., 2013; Schönauer et al., 2017). The integration of AI into our lives has only just begun, and we hope that its integration into our science will be mutually beneficial.

We have only scratched the surface in terms of the potential contributions of AI to decision making and decision making to AI, but it is clear that there are many avenues by which the two can improve one another. By focusing on cognitive models of decision behavior, we can both analyze the behavior of neural networks and more readily

understand human behavior. This creates more complete explanations of behavior across natural and artificial systems, in turn leading to richer theories of decision making that were previously unknown or impossible to test. By leveraging the predictive power of AI and machine learning in ways like those we have outlined here, we suspect the field of decision making—and psychology and cognitive science as a whole—will grow in new and unexpected ways.

References

Adami, C., Schossau, J., & Hintze, A. (2016). Evolutionary game theory using agent-based methods. *Physics of Life Reviews*, *19*, 1–26. https://doi.org/10.1016/j.plrev.2016.08.015

Agrawal, A. K., McHale, J., & Oettl, A. (2023). *Artificial intelligence and scientific discovery: A model of prioritized search* (Tech. Rep.). National Bureau of Economic Research. https://doi.org/10.1016/j.respol.2024.104989

Anderson, J. R. (1990a). *Cognitive psychology and its implications*. W.H. Freeman/Times Books/Henry Holt.

Anderson, J. R. (1990b). *The adaptive character of thought*. Psychology Press.

Ashton, R. H., & Ashton, A. H. (1995). *Judgment and decision-making research in accounting and auditing*. Cambridge University Press.

Ay, N., & Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems*, *11*(1), 17–41. https://doi.org/10.1142/S0219525908001465

Back, T. (1996). *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press.

Banzhaf, W., Beslon, G., Christensen, S., Foster, J. A., Képès, F., Lefort, V., Miller, J. F., Radman, M., & Ramsden, J. J. (2006). From artificial evolution to computational evolution: A research agenda. *Nature Reviews Genetics*, *7*(9), 729–735. https://doi.org/10.1038/nrg1921

Bateson, W., & Mendel, G. (2013). *Mendel's principles of heredity*. Courier Corporation.

Bello, P. (2012). Cognitive foundations for a computational theory of mindreading. *Advances in Cognitive Systems*, *1*(1–6), 59–72. http://www.cogsys.org/journal/volume1/

Bhatia, S. (2023). Inductive reasoning in minds and machines. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000446

Bhatia, S., & Mullett, T. L. (2018). Similarity and decision time in preferential choice. *Quarterly Journal of Experimental Psychology*, *71*(6), 1276–1280. https://doi.org/10.1177/1747021818763054

Blokpoel, M., Wareham, T., Haselager, P., Toni, I., & van Rooij, I. (2018). Deep analogical inference as the origin of hypotheses. *The Journal of Problem Solving*, *11*(1), Article 3. https://doi.org/10.7771/1932-6246.1197

Boelts, J., Lueckmann, J.-M., Gao, R., & Macke, J. H. (2022). Flexible and efficient simulation-based inference for models of decision-making. *Elife*, *11*, Article e77220. https://doi.org/10.7554/elife.77220

Bogacz, R. (2007). Optimal decision-making theories: Linking neurobiology with behaviour. *Trends in Cognitive Sciences*, *11*(3), 118–125. https://doi.org/10.1016/j.tics.2006.12.006

Bohm, C., Kirkpatrick, D., Cao, V., & Adami, C. (2022). Information fragmentation, encryption and information flow in complex biological networks. *Entropy*, *24*(5), Article 735. https://doi.org/10.3390/e24050735

Brighton, H., & Gigerenzer, G. (2012). Are rational actor models "rational" outside small worlds. In S. Okasha & K. Binmore (Eds.), *Evolution and rationality: Decisions, co-operation, and strategic behavior* (pp. 84–109). Cambridge University Press.

Bryson, J. J. (2005). Modular representations of cognitive phenomena in AI, psychology and neuroscience. In D. N. Davis (Ed.), *Visions of mind: Architectures for cognition and affect* (pp. 66–89). IGI Global. https://doi.org/10.4018/978-1-59140-482-8.ch004

Busemeyer, 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

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 133–154). John Wiley & Sons. https://doi.org/10.1002/9780470752937.ch7

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, *14*(3), 253–262. https://doi.org/10.1037/1040-3590.14.3.253

Carter, O., Hohwy, J., Van Boxtel, J., Lamme, V., Block, N., Koch, C., & Tsuchiya, N. (2018). Conscious machines: Defining questions. *Science*, *359*(6374), Article 400. https://doi.org/10.1126/science.aar4163

Cassenti, D. N., Veksler, V. D., & Ritter, F. E. (2022). *Editor's review and introduction: Cognition-inspired artificial intelligence* (Vol. 14, No. 4). Wiley Online Library.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional

form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty*, *52*(3), 233–254. https://doi.org/10.1007/s11166-016-9242-y

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*(4), 887–905. https://doi.org/10.1162/neco.2009.02-09-959

Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision. *Current Biology*, *24*(18), R921–R929. https://doi.org/10.1016/j.cub.2014.08.026

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117

Cybenko, G. V. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*, 303–314. https://doi.org/10.1007/BF02551274

Dehaene, S., Lau, H., & Kouider, S. (2021). What is consciousness, and could machines have it? In J. von Braun, M. S. Archer, G. M. Reichberg, & M. S. Sorondo (Eds.), *Robotics, AI, and humanity: Science, ethics, and policy* (pp. 43–56). Springer. https://doi.org/10.1007/978-3-030-54173-6_4

Descartes, R. (1901). *Discourse on method, and metaphysical meditations* (G. B. Rawlings, Trans.). W. Scott. https://doi.org/10.5962/bhl.title.44504

Drugowitsch, J., Mendonça, A. G., Mainen, Z. F., & Pouget, A. (2019). Learning optimal decisions with confidence. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(49), 24872–24880. https://doi.org/10.1073/pnas.1906787116

Dyer, F. C. (2002). The biology of the dance language. *Annual Review of Entomology*, *47*(1), 917–949. https://doi.org/10.1146/annurev.ento.47.091201.145306

Ellefsen, K. O., Mouret, J.-B., & Clune, J. (2015). Neural modularity helps organisms evolve to learn new skills without forgetting old skills. *PLOS Computational Biology*, *11*(4), Article e1004128. https://doi.org/10.1371/journal.pcbi.1004128

Elsemüller, L., Schnuerch, M., Bürkner, P.-C., & Radev, S. T. (2023). *A deep learning method for comparing Bayesian hierarchical models*. arXiv preprint. https://doi.org/10.48550/arXiv.2301.11873

Fang, J., Schooler, L., & Shenghua, L. (2023). Machine learning strategy identification: A paradigm to uncover decision strategies with high fidelity. *Behavior Research Methods*, *55*, 263–284. https://doi.org/10.3758/s13428-022-01828-1

Fengler, A., Govindarajan, L. N., & Frank, M. J. (2020). Encoder-decoder neural architectures for fast amortized inference of cognitive process models. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 1859–1865). Cognitive Science Society.

Fific, M., & Buckmann, M. (2013). Stopping rule selection (SRS) theory applied to deferred decision making. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2273–2278). Cognitive Science Society.

Finkel, E. (2023). Consciousness hunt yields results but not clarity. *Science*, *380*(6652), 1309–1310. https://doi.org/10.1126/science.adj4498

Fleming, S., Frith, C., Goodale, M., Lau, H., LeDoux, J. E., Lee, A. L., Michel, M., Owen, A. M., Peters, M. A. K., & Slagter, H. A. (2023). *The integrated information theory of consciousness as pseudoscience*. PsyArXiv. https://doi.org/10.31234/osf.io/zsr78

Gigerenzer, G. (2023). Psychological AI: Designing algorithms informed by human psychology. *Perspectives on Psychological Science*. Advance online publication. https://doi.org/10.1177/17456916231180597

Gigerenzer, G., & Todd, P. M. (1999). *Fast and frugal heuristics: The adaptive toolbox*. Oxford University Press.

Goldsmith, T. H. (1990). Optimization, constraint, and history in the evolution of eyes. *The Quarterly Review of Biology*, *65*(3), 281–322. https://doi.org/10.1086/416840

Gould, J. L., Henerey, M., & MacLeod, M. C. (1970). Communication of direction by the honey bee: Review of previous work leads to experiments limiting olfactory cues to test the dance language hypothesis. *Science*, *169*(3945), 544–554. https://doi.org/10.1126/science.169.3945.544

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. https://doi.org/10.1111/tops.12142

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), Article eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Gutmann, M. U., & Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, *17*(125), 1–47. https://jmlr.org/papers/v17/15-017.html

Happel, B. L., & Murre, J. M. (1994). Design and evolution of modular neural network architectures. *Neural Networks*, *7*(6–7), 985–1004. https://doi.org/10.1016/S0893-6080(05)80155-8

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, *9*(2), 394–401. https://doi.org/10.3758/BF03196299

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 25–48). Springer.

Hintze, A. (2021). *The role weights play in catastrophic forgetting* [Conference session]. 2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI), Institute of Electrical and Electronics Engineers, Inc.

Hintze, A., & Adami, C. (2022). Neuroevolution gives rise to more focused information transfer compared to backpropagation in recurrent neural networks. *Neural Computing and Applications*. Advance online publication. https://doi.org/10.1007/s00521-022-08125-0

Hintze, A., & Adami, C. (2023). Detecting information relays in deep neural networks. *Entropy*, *25*(3), Article 401. https://doi.org/10.3390/e25030401

Hintze, A., Edlund, J. A., Olson, R. S., Knoester, D. B., Schossau, J., Albantakis, L., Tehrani-Saleh, A., Kvam, P., Sheneman, L., Goldsby, H., Bohm, C., & Adami, C. (2017). *Markov brains: A technical introduction*. ArXiv. https://doi.org/10.48550/arXiv.1709.05601

Hintze, A., Imam, Y., & Rönnegård, L. (2022). *Testing the efficiency of a genome-wide association study on a computational evolutionary model* [Conference session]. ALIFE 2022: The 2022 Conference on Artificial Life.

Hintze, A., Olson, R. S., Adami, C., & Hertwig, R. (2015). Risk sensitivity as an evolutionary adaptation. *Scientific Reports*, *5*(1), Article 8242. https://doi.org/10.1038/srep08242

Hintze, A., Schossau, J., & Bohm, C. (2019). The evolutionary buffet method. In W. Banzhaf, L. Spector, & L. Sheneman (Eds.), *Genetic programming theory and practice XVI* (pp. 17–36). Springer.

Holmes, W. R. (2015). A practical guide to the Probability Density Approximation (PDA) with improved implementation and error characterization. *Journal of Mathematical Psychology*, *68–69*, 13–24. https://doi.org/10.1016/j.jmp.2015.08.006

Hopfield, J. J., & Tank, D. W. (1985). "Neural" computation of decisions in optimization problems. *Biological Cybernetics*, *52*(3), 141–152. https://doi.org/10.1007/BF00339943

Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, *340*(6132), 639–642. https://doi.org/10.1126/science.1234330

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*(1), 1–32. https://doi.org/10.1037/a0034190

Josephson, J. R., & Josephson, S. G. (1996). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Kaber, D. B., Wright, M. C., Prinzel, L. J., III, & Clamann, M. P. (2005). Adaptive automation of human-machine system information-processing functions. *Human Factors*, *47*(4), 730–741. https://doi.org/10.1518/001872005775570989

Kaplan, M. F., & Schwartz, S. (2013). *Human judgment and decision processes in applied settings*. Academic Press.

Kenny, C., & Pahl, C. (2009). Intelligent and adaptive tutoring for active learning and training environments. *Interactive Learning Environments*, *17*(2), 181–195. https://doi.org/10.1080/10494820802090277

Kidziński, Ł., Ong, C., Mohanty, S. P., Hicks, J., Carroll, S., Zhou, B., Zeng, H., Wang, F., Lian, R., Tian, H., Jaśkowski, W., Andersen, G., Lykkebø, O. R., Toklu, N. E., Shyam, P., Srivastava, R. K., Kolesnikov, S., Hrinchuk, O., Pechenko, A., … Tian, H. (2020). Artificial intelligence for prosthetics: Challenge solutions. In S. Escalera & R. Herbrich (Eds.), *The NeurIPS'18 competition: From machine learning to intelligent conversations* (pp. 69–128). Springer.

Krogman, W. M. (1951). The scars of human evolution. *Scientific American*, *185*(6), 54–57. https://doi.org/10.1038/scientificamerican1251-54

Kuniyoshi, Y., Yorozu, Y., Ohmura, Y., Terada, K., Otani, T., Nagakubo, A., & Yamamoto, T. (2004). From humanoid embodiment to theory of mind. In F. Iida, R. Pfeifer, L. Steels, & Y. Kuniyoshi (Eds.), *Embodied artificial intelligence: International seminar, Dagstuhl Castle, Germany, July 7–11, 2003, revised papers* (pp. 202–218). Springer.

Kvam, P. D. (2019). A geometric framework for modeling dynamic decisions among arbitrarily many alternatives. *Journal of Mathematical Psychology*, *91*, 14–37. https://doi.org/10.1016/j.jmp.2019.03.001

Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review*, *127*(6), 1053–1078. https://doi.org/10.1037/rev0000215

Kvam, P. D., & Hintze, A. (2018). Rewards, risks, and reaching the right strategy: Evolutionary paths from heuristics to optimal decisions. *Evolutionary Behavioral Sciences*, *12*(3), 177–190. https://doi.org/10.1037/ebs0000115

Kvam, P. D., Hintze, A., Pleskac, T. J., & Pietraszewski, D. (2019). Computational evolution and ecologically rational decision making. In R. Hertwig, T. J. Pleskac, & T. Pachur (Eds.), *Taming Uncertainty* (pp. 285–304). MIT Press.

Kvam, P. D., Smith, C., Irving, L. H., & Sokratous, K. (2024). Improving the reliability and validity of the IAT with a dynamic model driven by associations. *Behavior Research Methods*, *56*(3), 2158–2193. https://doi.org/10.3758/s13428-023-02141-1

Kvam, P. D., & Turner, B. M. (2021). Reconciling similarity across models of continuous selections. *Psychological Review*, *128*(4), 766–786. https://doi.org/10.1037/rev0000296

Kwisthout, J., & Van Rooij, I. (2013). Bridging the gap between theory and practice of approximate bayesian inference. *Cognitive Systems Research*, *24*, 2–8. https://doi.org/10.1016/j.cogsys.2012.12.008

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., Cheney, N., Chrabaszcz, P., Cully, A., Doncieux, S., Dyer, F. C., Ellefsen, K. O., Feldt, R., Fischer, S., Forrest, S., … Yosinksi, J. (2018). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, *26*(2), 274–306. https://doi.org/10.1162/artl_a_00319

Liefooghe, B., & Van Maanen, L. (2023). Three levels at which the user's cognition can be represented in artificial intelligence. *Frontiers in Artificial Intelligence*, *5*, Article 1092053. https://doi.org/10.3389/frai.2022.1092053

Ling, S. H., & Lam, H. K. (2019). Evolutionary algorithms in health technologies. *Algorithms*, *12*(10), Article 202. https://doi.org/10.3390/a12100202

Lueckmann, J.-M., Bassetto, G., Karaletsos, T., & Macke, J. H. (2019). *Likelihood-free inference with emulator networks*. arXiv. https://doi.org/10.48550/arXiv.1805.09294

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnet (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 1–10). Curran Associates, Inc.

Ma, Y., Wang, Z., Yang, H., & Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: A survey. *IEEE/CAA Journal of Automatica Sinica*, *7*(2), 315–329. https://doi.org/10.1109/JAS.2020.1003021

Marr, D. (1982). *Computational investigation into the human representation and processing of visual information*. MIT Press.

Marstaller, L., Hintze, A., & Adami, C. (2013). The evolution of representation in simple cognitive networks. *Neural Computation*, *25*(8), 2079–2107. https://doi.org/10.1162/neco_a_00475

Matthews, M., Chowdhary, G., & Kieson, E. (2017). *Intent communication between autonomous vehicles and pedestrians*. arXiv preprint. https://doi.org/10.48550/arXiv.1708.07123

McCoy, J. T., Kroon, S., & Auret, L. (2018). Variational autoencoders for missing data imputation with application to a simulated milling circuit. *IFAC-PapersOnLine*, *51*(21), 141–146. https://doi.org/10.1016/j.ifacol.2018.09.406

McMullen, D. P., Hotson, G., Katyal, K. D., Wester, B. A., Fifer, M. S., McGee, T. G., Harris, A., Johannes, M. S., Vogelstein, R. J., Ravitz, A. D., Anderson, W. S., Thakor, N. V., & Crone, N. E. (2013). Demonstration of a semi-autonomous hybrid brain–machine interface using human intracranial EEG, eye tracking, and computer vision to control a robotic upper limb prosthetic. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(4), 784–796. https://doi.org/10.1109/tnsre.2013.2294685

McMullin, E. (2013). The virtues of a good theory. In S. Psillos & M. Curd (Eds.), *The Routledge companion to philosophy of science* (pp. 561–571). Routledge.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92. https://doi.org/10.1016/j.neuron.2015.09.039

Miikkulainen, R. (2021). Creative AI through evolutionary computation: Principles and examples. *SN Computer Science*, *2*(3), Article 163. https://doi.org/10.1007/s42979-021-00540-9

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., & Hodjat, B. (2019). Evolving deep neural networks. In R. Kozma, C. Alippi, Y. Choe, & F. C. Morabito (Eds.), *Artificial intelligence in the age of neural networks and brain computing* (pp. 293–312). Elsevier.

Miletić, S., Turner, B. M., Forstmann, B. U., & van Maanen, L. (2017). Parameter recovery for the leaky competing accumulator model. *Journal of Mathematical Psychology*, *76*(Pt. A), 25–50. https://doi.org/10.1016/j.jmp.2016.12.001

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Nguyen, T. N., & Gonzalez, C. (2022). Theory of mind from observation in cognitive models and humans. *Topics in Cognitive Science*, *14*(4), 665–686. https://doi.org/10.1111/tops.12553

Oliveira, A. L. (2022). A blueprint for conscious machines. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(23), Article e2205971119. https://doi.org/10.1073/pnas.2205971119

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20–48. https://doi.org/10.1016/j.jmp.2018.03.003

Palestro, J. J., Sederberg, P. B., Osth, A. F., Van Zandt, T., & Turner, B. M. (2018). *Likelihood-free methods for cognitive science* (1st ed.). Springer.

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

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1016/j.tics.2008.02.010

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In J. Dy & A. Krause (Eds.), *International conference on machine learning* (pp. 4218–4227). Curran Associates, Inc.

Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., & Bürkner, P.-C. (2021). Amortized Bayesian model comparison with evidential deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 4903–4917. https://doi.org/10.1109/TNNLS.2021.3124052

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020). *BayesFlow: Learning complex stochastic models with invertible neural networks*. arXiv. https://doi.org/10.48550/arXiv.2003.06281

Radev, S. T., Mertens, U. K., Voss, A., & Köthe, U. (2020). Towards end-to-end likelihood-free inference with convolutional neural networks. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 23–43. https://doi.org/10.1111/bmsp.12159

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. (2018). Decision making on spatially continuous scales. *Psychological Review*, *125*(6), 888–935. https://doi.org/10.1037/rev0000117

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

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). Association for Computing Machinery.

Rmus, M., Pan, T.-F., Xia, L., & Collins, A. G. (2023). *Artificial neural networks for model identification and parameter estimation in computational cognitive models*. bioRxiv. https://doi.org/10.1101/2023.09.14.557793

Rohrbaugh, C. C., Shanteau, J., & Hall, B. (1999). Context, process, and experience: Research on applied judgment and decision making. In F. T. Durso, R. S. Nickerson, S. T. Dumais, S. Lewandowsky, & T. J. Perfect (Eds.), *Handbook of applied cognition* (pp. 115–139). John Wiley & Sons.

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education.

Santos, L. R., & Rosati, A. G. (2015). The evolutionary roots of human decision making. *Annual Review of Psychology*, *66*, 321–347. https://doi.org/10.1146/annurev-psych-010814-015310

Savage, L. J. (1954). *The foundations of statistics*. Wiley.

Scassellati, B. M. (2001). *Foundations for a theory of mind for a humanoid robot* [Unpublished doctoral dissertation]. Massachusetts Institute of Technology.

Schönauer, M., Alizadeh, S., Jamalabadi, H., Abraham, A., Pawlizki, A., & Gais, S. (2017). Decoding material-specific memory reprocessing during sleep in humans. *Nature Communications*, *8*(1), Article 15404. https://doi.org/10.1038/ncomms15404

Schossau, J., & Hintze, A. (2020). Small implementation differences can have large effects on evolvability: "Even the largest avalanche is triggered by small things"—Vernor vinge. In W. Banzhaf, B. H. C. Cheng, K. Deb, K. E. Holekamp, R. E. Lenski, C. Ofria, R. T. Pennock, W. F. Punch, & D. J. Whittaker (Eds.), *Evolution in action: Past, present and future: A festschrift in honor of Erik D. Goodman* (pp. 423–434). Springer.

Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., & Radev, S. T. (2023). Neural superstatistics for bayesian estimation of dynamic cognitive models. *Scientific Reports*, *13*(1), Article 13778. https://doi.org/10.1038/s41598-023-40278-3

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., Hug, B., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*(6), 632–654. https://doi.org/10.1002/tea.20311

Sclar, M., Neubig, G., & Bisk, Y. (2022). Symmetric machine theory of mind. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *International conference on machine learning* (pp. 19450–19466). Curran Associates, Inc.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shafi, M. M., Westover, M. B., Fox, M. D., & Pascual-Leone, A. (2012). Exploration and modulation of brain network interactions with noninvasive brain stimulation in combination with neuroimaging. *European Journal of Neuroscience*, *35*(6), 805–825. https://doi.org/10.1111/j.1460-9568.2012.08035.x

Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in Cognitive Science*, *2*(4), 736–750. https://doi.org/10.1111/j.1756-8765.2010.01092.x

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai,

M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Simon, H. A. (1981). Studying human intelligence by creating artificial intelligence: When considered as a physical symbol system, the human brain can be fruitfully studied by computer simulation of its processes. *American Scientist*, *69*(3), 300–309. https://www.jstor.org/stable/27850429

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*, 1–20. https://doi.org/10.1146/annurev.ps.41.020190.000245

Sokratous, K., Fitch, A. K., & Kvam, P. D. (2023). How to ask twenty questions and win: Machine learning tools for assessing preferences from small samples of willingness-to-pay prices. *Journal of Choice Modelling*, *48*, Article 100418. https://doi.org/10.1016/j.jocm.2023.100418

Stevenson, M. K., Busemeyer, J. R., & Naylor, J. C. (1990). *Judgment and decision-making theory*. Consulting Psychologists Press.

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, *22*(9), 1503–1511. https://doi.org/10.1038/s41593-019-0453-9

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, *28*(2), 454–475. https://doi.org/10.3758/s13423-020-01825-5

Tehrani-Saleh, A., & Adami, C. (2020). Can transfer entropy infer information flow in neuronal circuits for cognitive processing? *Entropy*, *22*(4), Article 385. https://doi.org/10.3390/e22040385

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*, Article 42. https://doi.org/10.1186/1471-2202-5-42

Towne, W. F., & Gould, J. L. (1988). The spatial precision of the honey bees' dance communication. *Journal of Insect Behavior*, *1*, 129–155. https://doi.org/10.1007/BF01052234

Townsend, J. T., & Busemeyer, J. R. (2014). Approach-avoidance: Return to dynamic decision behavior. In C. Izawa (Ed.), *Cognitive processes the Tulane flowerree symposia on cognition* (pp. 107–133). Psychology Press.

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology*, *56*(5), 375–385. https://doi.org/10.1016/j.jmp.2012.06.004

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250. https://doi.org/10.3758/s13423-013-0530-0

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

van Ravenzwaaij, D., van der Maas, H. L., & Wagenmakers, E.-J. (2012). Optimal decision making in neural inhibition models. *Psychological Review*, *119*(1), 201–215. https://doi.org/10.1037/a0026275

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–85. https://doi.org/10.1146/annurev.psych.60.110707.163633

Yakovenko, N., Cao, L., Raffel, C., & Fan, J. (2016). Poker-CNN: A pattern learning strategy for making draws and bets in poker games using convolutional networks. *Proceedings of the AAAI conference on artificial intelligence*, *30*(1). https://doi.org/10.1609/aaai.v30i1.10013

Yildirim, I., & Paul, L. (2023). *From task structures to world models: What do LLMs know?* arXiv preprint. https://doi.org/10.48550/arXiv.2310.04276

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In G. A. Papadopoulos (Ed.), *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 307–317). Association for Computing Machinery.

Zhang, W., Gao, B., Tang, J., Yao, P., Yu, S., Chang, M.-F., Yoo, H.-J., Qian, H., & Wu, H. (2020). Neuro-inspired computing chips. *Nature Electronics*, *3*(7), 371–382. https://doi.org/10.1038/s41928-020-0435-7

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 295–305). Association for Computing Machinery.

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., & Goyal, H. (2018). Opening the black box of neural networks: Methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine*, *6*(11), Article 216. https://doi.org/10.21037/atm.2018.05.32

Zhou, D.-X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, *48*(2), 787–794. https://doi.org/10.1016/j.acha.2019.06.004

Received May 25, 2023
Revision received February 27, 2024
Accepted March 3, 2024 ■