Towards machines that understand people

Howes, Andrew; Jokinen, Jussi P. P.; Oulasvirta, Antti

[Link to publication on Research at Birmingham portal](#)

Towards machines that understand people

Andrew Howes[1] | **Jussi P. P. Jokinen**[2] | **Antti Oulasvirta**[3]

[1]University of Exeter, Exeter, UK

[2]University of Jyväskylä, Jyväskylä, Finland

[3]Aalto University, Espoo, Finland

Correspondence
Andrew Howes, University of Exeter, Exeter, UK.
Email: a.howes@bham.ac.uk

Abstract

The ability to estimate the state of a human partner is an insufficient basis on which to build cooperative agents. Also needed is an ability to predict how people adapt their behavior in response to an agent's actions. We propose a new approach based on computational rationality, which models humans based on the idea that predictions can be derived by calculating policies that are approximately optimal given human-like bounds. Computational rationality brings together reinforcement learning and cognitive modeling in pursuit of this goal, facilitating machine understanding of humans.

INTRODUCTION

Building machines that understand people is a key goal for research on cooperative AI (Lake et al. 2017; Dafoe et al. 2021; Hadfield-Menell et al. 2016; Anderson, Boyle, and Reiser 1985; Batmaz et al. 2019), with applications in adaptive interfaces, automated vehicles, human–robot interaction, tutoring, decision support, and design. How, for example, is it possible for a machine to understand why Tina took the long route to work (Figure 1)? The answer might concern her utility function (her preferences) or her capacities (including memory and attention). She might have taken route C, for example, because she prefers the scenery or because it has fewer intersections and, therefore, implies lower cognitive workload. Arguably, the key challenge to solving this type of problem is in the complexity of the latent psychological processes that underpin behavior. This is a problem of identifiability; any observed human behavior may have arisen due to one of a number of possible underlying processes. Figuring out which latent processes cause which events is a hard problem at the cutting edge of Artificial Intelligence and Cognitive Science research (e.g., Dafoe et al. 2021; Gershman and Daw 2017; Jara-Ettinger 2019; Lake et al. 2017; Russell 2019).

A related problem concerns how to predict behavior in unseen, or counterfactual circumstances, that is by answering "what if?" questions. In Figure 1, for example, to predict whether it is worth suggesting route B, a cooperative agent might conduct a number of "what if?" simulations in order to determine the probable outcome of this intervention. The simulations might be conducted under a distribution of assumptions about Tina's subjective utilities and capacities; in other words, planning could be assisted by asking "what if?" questions of a *model*. In the current article, we argue that answering "what-if?" and "why?" questions about human behavior requires a very specific type of model known as a *Computationally Rational* model. Their goal is to capture the extraordinary capacity of the human mind to adapt behavior in service of an individual's utilities, capabilities, and the environment. They are information processing models of cognition in which predictions are made by a decision policy that is optimally adapted to the internal factors that limit human behavior (Gershman, Horvitz, and Tenenbaum 2015; Lewis, Howes, and Singh 2014; Lieder and Griffiths 2019). The limits (or bounds) include individual utility functions, capacities, and experiences. The limits are stated as a unified theory of the mechanisms of cognition and are not
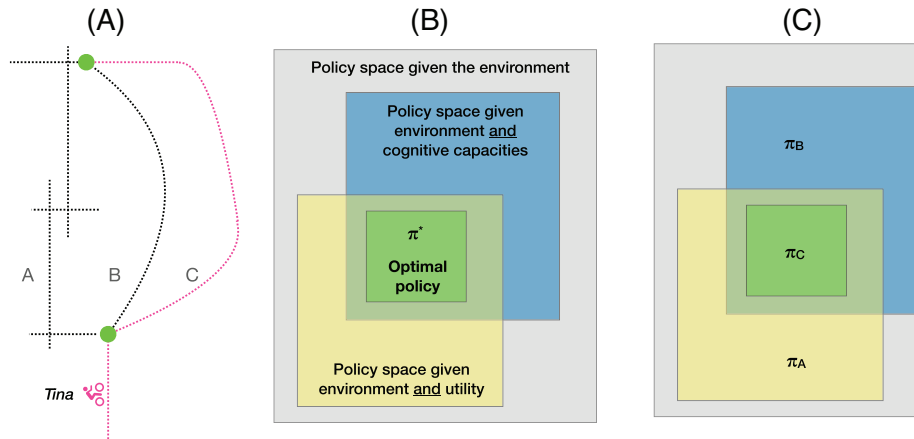
FIGURE 1 Panel (A): *Why* did Tina take the long route C? Was it because of the scenery (a subjective utility) or because it had lower workload (a capacity limit)? In order to assist a human partner, an AI assistant must be able to (1) reason about causes of behavior by using a model to answer *why?* questions, and (2) predict the consequences of its actions, by using the same model to answer *what if?* questions. For example, would advising Tina to take route A, through a city with multiple interchanges, or B be annoying or useful? **Computational rationality** models human behavior as optimization of expected utility under bounds (possible causes of behavior). Human behavior is predicted by the optimal policy π^* given the environment, cognitive capacities and utility of the individual. Panel (B) illustrates how bounds limit the policy space and thereby permit the prediction of human behavior. Given a utility function (a model of human subjective utility), computational rationality considers the space determined jointly by the environment and human capacities (intersection of yellow and blue squares). The optimal subset of these policies (green square) π^* provides a basis for cooperative AI to understand people. Panel (c): Tina's available behaviors are illustrated as consequences of different policies and therefore different causal bounds allowing why questions to be answered. Why did not Tina take route A? Answer: because it is only a consequence of a suboptimal π_A with respect to capacity limits. Why not B? Because B is a consequence of a suboptimal policy with respect to utility. Why C? Because C is a consequence of a bounded optimal policy with respect to bounds on environment, capacity limits, and utility. What if any of these bounds change, or are changed by the cooperative AI? Then the analysis of the policy space must be rerun and the implications of the changes for behavior calculated.

merely a weighted set of factors. Modeling these limits, and the ability to calculate their implications for behavior, must be a core challenge for research on cooperative AI.

Computational rationality explains human behavior by establishing a causal connection between subjective utility, capacities, and experience on the one-hand and behavior on the other. Computationally rational models achieve this by finding a policy that maximizes utility under bounds. Deployed in cooperative agents, both "what if" and "why?" questions can be answered by iteratively posing problems to the model and then running it to determine, after factoring in adaptation, the predicted outcome.

Computational rationality brings together a number of ideas from cognitive science, including ideas from cognitive architectures (Newell 1990; Meyer and Kieras 1997; Anderson 2014; Laird 2019; Forbus and Hinrich 2017), ecological rationality (Luan, Reb, and Gigerenzer 2019), and Bayesian rationality (Oaksford and Chater 2007; Anderson 1990). What computational rationality adds is that predictions are made by calculating the bounded optimal policy in a given environment. It seeks explanations of behavior as optimal adaptations to *both* ecological and cognitive bounds (including intentions) (Howes, Lewis, and Vera 2009; Lewis, Howes, and Singh

2014). It covers not only information processing capacities but also the way that their use is adaptive.

Computationally rational models can explain *why* behavior differs among individuals and conditions, as opposed to just describing them. In psychology, it has been used to explain how behavior adapts to a number of cognitive limits, including the limits of memory (Howes et al. 2016), perception (Geisler 2011), and motor system (Tseng and Howes 2015). It has also been used to explain how behavior adapts to the environment, visible, for example, in multitasking strategies that adapt to switch costs of the task environment (Brumby, Salvucci, and Howes 2007), the probability of error (Quinn and Zhai 2016), and information gains in documents (Duggan and Payne 2009).

As a model-based approach, computational rationality is a departure from the repertoire of model-free machine learning methods for cooperative AI. Arguably, the latent states and processes that are characteristic of human cognition are not captured easily by model-free methods that learn from patterns in observed behavior. In probabilistic methods, computationally rational models can serve as a strong prior to help AI reason better about people, through counterfactual experiments that reference psychologically meaningful constructs concerning subjective utilities, goals, cognitive limits, and experience. Other

researchers have gone as far as arguing that progress in AI will curtail without better use of models (Pearl 2021; Lake et al. 2017). Given the intrinsic difficulty of the problem, the problem of interacting with humans could well be one of the problems where this is true. However, while there is an abundance of work on how AI agents can model other artificial agents (Albrecht and Stone 2018), there is much less on how to model humans.

This paper reviews assumptions and progress in computational rationality, an area that has developed rapidly during the last decade thanks to a convergence of research in machine learning and cognitive science (Dayan and Daw 2008; Lewis, Howes, and Singh 2014; Gershman, Horvitz, and Tenenbaum 2015; McClelland and Botvinick; Lake et al. 2017; Lieder and Griffiths 2019). One core insight has been that the adaptive problems faced by people can be defined as reinforcement learning (RL) problems and solved accordingly. The vast range of modern RL (and deep-RL) methods provide the means to derive approximately optimal policies given formal specifications of the problems faced by people. In addition, deep neural networks have made contributions to understanding not only computing optimal policies for complex state-action spaces, but also in modeling human bounds such as in perception. Thanks to this convergence, computational rationality is pushing to the forefront in a range of applied problems where people demonstrate adaptive behavior, particularly in human–computer interaction (HCI) and robotics (Chen et al. 2015, 2017; Jokinen et al. 2020; Chen et al. 2021; Hadfield-Menell et al. 2016; Fisac et al. 2020; Oulasvirta, Jokinen, and Howes 2022).

The assumption that human behavior is boundedly optimal is controversial in cognitive science (Rahnev and Denison 2018). Some researchers claim evidence against it, and even support the idea that people are often irrational (Bowers and Davis 2012; Colombo, Elkin, and Hartmann 2020). However, there is evidence that people appear optimal when the bounds imposed by computation or environment are taken into account (Swets, Tanner Jr, and Birdsall 1961; Maloney and Mamassian 2009; Geisler 2011; Körding and Wolpert 2006; Oaksford and Chater 2007; Gigerenzer 2018; Hahn and Warren 2009; Howes et al. 2016; Bahrami et al. 2010; Howes, Lewis, and Vera 2009; Juechems et al. 2021; Callaway, Antonio, and Tom 2020; Chen et al. 2015; Lewis, Howes, and Singh 2014). There is also emerging evidence in social interaction. People assume that other people are bounded optimal in order to understand their behavior (Johnson and Rips 2015; Bridgers, Jara-Ettinger, and Gweon 2020). For the purposes of cooperative AI, the question is not whether people are able to optimize behavior, but whether behavior can be predicted by an optimization algorithm. Indeed, a number of approaches to cooperative AI now assume that human collaborators behave rationally and calculate behavioral predictions accordingly (Consul et al. 2021; Hadfield-Menell et al. 2016; Fisac et al. 2020).

In the rest of the paper, we first explain the background to computational rationality, paying particular attention to the adaptive nature of cognition. We then outline the general structure of an approach to cooperative AI, focusing on how models can be used to answer "what if?" questions, including the problems of intervention and inference. Finally, we turn to outstanding challenges and propose an agenda for research.

UNDERSTANDING HUMAN ADAPTATION AS COMPUTATIONAL RATIONALITY

Computationally, rational models view cognition as a bounded optimization problem (Lewis, Howes, and Singh 2014; Lieder and Griffiths 2019). *Bounded optimality* (Russell and Subramanian 1994), is distinguished from *bounded rationality*, where, in the latter, the role of optimization in understanding behavior is rejected (Chase, Hertwig, and Gigerenzer 1998). An agent is bounded optimal if its program is a solution to the optimization problem presented by its utility function (objective), capacity limits (architecture), and the task environment. When a model of a human and the environment defines a bounded optimality problem, the solution is a program P^* (Lewis, Howes, and Singh 2014)—a policy π^* (see Equation 1 in Box 1). The program simulates the user's strategy, which determines the choice of actions. P^* is adapted through optimization to the bounds imposed by the tuple, possibly using machine learning. In Box 1, Equation (1), this assumption is highlighted by stating that the optimal policy π^* is a function of the agent's subjective utility, capacity limits, and experience of the environment (its history).

Computational rationality is a departure from the earlier idea that human behavior can be predicted by considering only the environment, and, "not the assumed structure of the mind" (Anderson 1991). It is also a departure from the idea that behavior is shaped to external rewards. Internal bounds matter. Different assumptions about the bounds on adaptation are illustrated in Figure 1. The largest, outer-box, represents the space of policies defined by the environment. There is an optimal space of policies given just the environment (Anderson 1991) but the space of policies given the environment *and* the agent's capacities is a different, possibly overlapping, space. The optimal space of policies given this space is the smaller subset, π^*.

A computationally rational model demands the specification of an optimization problem in terms of human subjective utilities, capacities, and environment (Lewis,

Howes, and Singh 2014). If a cooperative AI is to explain human behavior, then it must do so in terms of a model of these components of the problem and a derived optimal policy. In other words, cooperative AI needs a computational model of the human that is defined in terms of an objective function representing human subjective utility, a model of human capacities (memory, perception, reasoning, etc.), and a model of how humans experience the task environment over a lifetime.

In addition, to be utilized in cooperative AI, the general human model needs to be "fit" to each individual collaborator. People vary widely along all three of the key dimensions. For example, in subjective utility, while most people are risk-averse, there is considerable variation (Mata et al. 2018), with some individuals even being risk-seeking in the domain of gains. Similarly, internal bounds vary. Humans form intentions in the context of these preferences and cooperative AI must be informed by a model of human preferences if it is to be successful (Russell 2019). While admitting to the possibility of variation in subjective weightings, computational rationality commits to a normative model of subjective utility, which is a requirement of the rationality assumption. Rather than appealing to "biases" in choice, a computationally rational account explains behavior as an emergent consequence of resource limits rather than as a consequence of "irrational" policies (Howes et al. 2016; Juechems et al. 2021). In Howes et al. (2016), for example, contextual effects on choice are explained in terms of the agent's uncertainty about expected value, rather than in terms of distortions of utility. Similarly, Juechems et al. (2021) show that apparent distortions of value and probability are consequences of optimal decision making given the assumption that humans operate with finite computational precision (Bhui, Lai, and Gershman (2021) provide a review).

Humans also vary in capabilities and experience of the task environment. While most people have a working memory capacity between 3 and 5 items (Cowan 2010), for example, some remember fewer, and some more, items (Conway, Jarrold, and Miyake 2008). And, of course, experiences also vary, partly through the laws of chance, but sometimes also due to whether statistical information is acquired through experience or through description (Rakow and Newell 2010). As regards, the task environment, there is evidence that humans are adapted to the task environment with extraordinary subtlety (Anderson 1990; Oaksford and Chater 2007). Simon (1969) pointed out that it is impossible to understand human behavior without understanding that it is adapted to both the structure of the environment and cognitive limitations. Computational rationality provides a formal framework for investigating both as constraints on human behavior. We believe that it is impossible to explain human

behavior without taking utility, capabilities, and experienced environment into account simultaneously.

In sum, computational rationality explains individual behavior (Howes, Lewis, and Vera 2009) despite the difficulties associated with diversity by providing the means to establish a causal link from subjective utilities, capacities, and experience, via the calculation of a bounded optimal policy, to observed behavior. In Box 1, the model is fit to individuals via parameter inference, which is carried out in Equation (2) and results in a set of parameters θ^* that govern the functioning of the model.

In cognitive science, it has been shown that explanations for some puzzling aspects of human behavior have sometimes been misattributed amongst these three components of the bounded optimization problem. Kahneman (Kahneman and Tversky 1979), for example, attributes a number of human behaviors to distorted utility functions when these should, instead, be attributed to the structure of the task environment (Gigerenzer 2018). Similarly, contextual effects on choice have been attributed to irrationality (Ariely 2009) when, they are more likely caused by rational adaptation to information provided by the context about the structure of the environment (Howes et al. 2016) and the way in which information is presented (Rakow and Newell 2010).

Coin flipping illustrates the issue (Hahn and Warren 2009). Consider a setting where a fair coin is thrown n times and we are interested in the probability of sequences with certain subsequences of length k where $k < n$. People have been shown to, *correctly*, believe that under these conditions, (e.g., $n = 4$ and $k = 3$) HHT (heads, heads, tails) will occur more frequently than HHH. This result is surprising to those who ignore the task environment and incorrectly calculate that the probability of HHH = $0.5 \times 0.5 \times 0.5$ = the probability of HHT. The correct calculation requires counting the relative frequencies of each sequence within sequences of length 4. The relative frequency of encountering at least one HHT is 4/16 = 0.25 whereas that of encountering HHH is 3/16 = 0.19. The reason that HHH occurs in fewer sequences of length $n = 4$ is that two of the occurrences overlap in a single sequence. It is wrong to cite the fact that a person believes that they will experience HHT more often than HHH in a sequence as evidence for a biased utility function. Hahn (Hahn and Warren 2009) reached this breakthrough with a mathematical analysis of the problem. A computationally rational model embedded in a cooperative AI system would reach the same conclusion.

Almost always, of course, behavior is attributable to some mix of all three components of the adaptation problem. When people perform more than one task at the same time, performance of the tasks can interfere; one or both tasks may be slowed, or the rate of errors may increase

FIGURE 2 Illustrations of two alternative Markovian definitions of the computationally rational problems faced by humans. Left: An agent operating directly in the external environment. Right: an agent operating via its internal environment (Singh, Lewis, and Barto 2009) and yoked to the external environment by stimulus and response. When the latter is used to model humans, cognitive bounds are modeled as functions bounding observations (o), reward (r), and internal action (a). Here, the agent operates in an internal environment; it observes percepts, motor system, memory, fatigue, pain, stress, and so forth. Stimuli are perceived and represented in the internal environment and responses are caused by changes in the internal environment. The reward is a *subjective* utility function capturing the agents preferences.
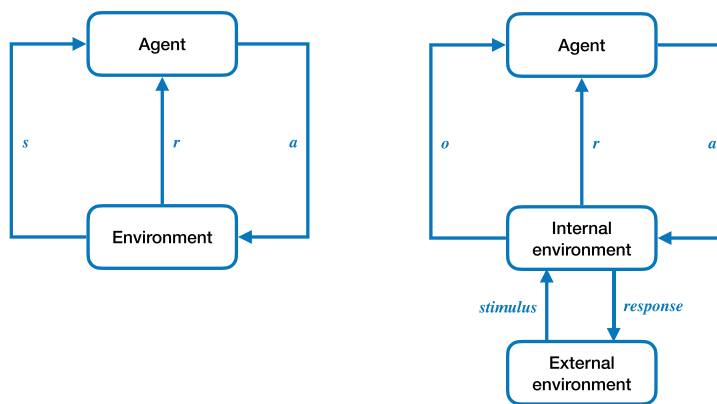
(Welford 1952). In addition, people can make strategic choices about which task to prioritize and whether to prioritize speed or accuracy (Meyer and Kieras 1997). The fact that task performance is influenced by both capacity limits and strategic concerns makes it difficult for an observer to tease out the contribution of each. A driver-assist AI system that attempts to infer whether a driver is (a) swerving to avoid a pothole, or (b) drifting due to simultaneously driving and browsing the web, must tease apart the causal roles of goals (leading to swerving) and resource limits combined with a unhealthy appetite for risk (leading to drifting) in multitask performance. It is impossible to understand human behavior without understanding that it is adapted to both the structure of the environment and the structure of cognition (Simon 1969). This is an important lesson for cooperative AI and one that raises questions that may be answered with computational rationality, which provides a framework for attributing environmental and cognitive causes to behavioral phenomena.

Recent evidence has started to speak for a human disposition to attribute rationality to others and thereby determine causes for behavior. The "naive utility calculus," or the ability to infer causes behind the observed behavior of others by assuming that they are bounded rational, seems to be an essential feature of human social interactions and theory of mind (Jara-Ettinger, Schulz, and Tenenbaum 2020). Importantly, the computational principles behind this ability seem to imply that humans are able to process this inference problem by presenting it as an inverse utility maximization or rational planning problem (Baker et al. 2017; Jara-Ettinger, Schulz, and Tenenbaum 2020). Computational rationality is, therefore, a plausible candidate for both simulating how humans infer latent states of each other, as well as for how a computational agent might implement a computational theory of mind, granting it the ability to better understand its users.

COMPUTATIONALLY RATIONAL MODELS OF COGNITION

In many, though not all, computationally rational models, the optimal policy is generated by defining the bounded optimality problem as a Markovian decision problem (MDP) and then solving it using classical planning algorithms (Dayan and Daw 2008; Daw 2014; Littman 2015; Lieder and Griffiths 2019). In addition to MDP, the approaches include partially observable Markov decision problems (POMDPs), meta-level MDPs, belief-MDPs, semi-MDPs, and multi-agent MDPs. All of these problems share the fact that they can be used to precisely capture a sequential bounded optimalisation problem in which goals are expressed as cumulative reward maximization (Bellman 1966; Sutton and Barto 2018). The definition of a MDP sets aside the mechanisms by which the optimization problem might be solved in favor of a precise formalization of the problem itself.

MDPs offer the flexibility needed to model hypotheses about human cognition, including hypotheses concerning bounds, hierarchies, reward-generating mechanisms, and so forth (see Figure 2). They have been used to model a large range of human decision problems (Daw 2014; Dayan and Daw 2008; Callaway, Rangel, and Griffiths 2021; Rieskamp and Otto 2006; Jara-Ettinger 2019; Spektor et al. 2019; Chen, Chang, and Howes 2021; Chen et al. 2017, 2015; Lieder and Griffiths 2019; Callaway, Rangel, and Griffiths 2021; Callaway et al. 2022), for example, prediction learning, where people learn what to expect will happen after a visual cue (Dayan and Daw 2008). Here, bounds imposed by the statistical structure of the task environment can be captured in the state transition probabilities. The formalism allows modeling internal bounds as well, such as bounds in the observation function (o in Figure 2) and transitions of the internal environment.

Computational rationality allows modeling cognition as an internal environment that is acted on by a controller. While in standard uses of Markovian formalisms, the agent interacts with an external environment, in some computationally rational models, an *agent* interacts with an *internal* environment via observations and actions and a yoked *external* environment via stimuli and responses. To our knowledge, the first extension of Markovian models to agents with internal environments was proposed by Barto and colleagues (Barto, Singh, and Chentanez 2004) and later Singh, Lewis, and Barto (2009); however, the first applications in complex human tasks emerged only a decade later (Chen et al. 2015). See Oulasvirta, Jokinen, and Howes (2022) for a review.

MDPs have also been used to model how people reason about the behavior of others (Jara-Ettinger 2019). They have also had particular success in modeling human visual search (Rao 2010; Acharya et al. 2017), attention allocation in simple choice tasks (Callaway, Rangel, and Griffiths 2021), and risky decision making (Chen et al. 2017; Chen and Howes 2012; Lieder, Krueger, and Griffiths 2017; Chen, Chang, and Howes 2021). For example, in Callaway, Rangel, and Griffiths (2021)'s meta-level Markovian model of attention allocation, beliefs are posterior distributions over choice item values and the reward function includes costs of computation. Even complex human behavior, such as multitasking, can be modeled by assuming a hierarchical organisation of multiple interconnected Markovian models (Jokinen et al. 2021; Jokinen, Kujala, and Oulasvirta 2020). MDPs are also being used to investigate human motivation, theories of which can be formulated as theories of subjective utility ($\mathcal{U}$ in Box 1) (Biehl et al. 2018). The internal reward framework proposed by Singh et al. (2010) is consistent with this view.

Box 1: Cooperative AI with computational rationality

The problem of cooperative AI is to maximize a value function $\mathcal{V}$ that summarizes some objective goodness of the interaction, while a human partner attempts to maximize their *own* subjective utility $\mathcal{U}$. The distinction between $\mathcal{V}$ and $\mathcal{U}$ allows modeling situations where the two objectives are not the same. Both $\mathcal{V}$ and $\mathcal{U}$ are defined as mappings from a history $h \in \mathcal{H}$ to a scalar value: $\mathcal{V} : \mathcal{H} \to \mathbb{R}$ and $\mathcal{U} : \mathcal{H} \to \mathbb{R}$, where $\mathcal{H}$ is the set of all possible histories.

A computationally rational model of the user: A task history $h \in \mathcal{H}$ is generated from the model M, following a policy π with parameters θ of the mechanisms of human cognition, in a task environment ϕ. A computationally rational agent (Lewis, Howes,

---

and Singh 2014) deploys a bounded optimal policy π^* that maximizes the utility of the task:

$$\pi^* = \arg\max_{\pi} \mathcal{U}(h), \qquad (1)$$

where $h \sim M(\pi, \theta, \phi)$.

Why? questions: Parameter inference determines the most plausible set of parameters θ^* by maximizing the likelihood of observed human data y:

$$\theta^* = \arg\max_{\theta} p(y \mid M) \qquad (2)$$

In practice, computationally rational models require a likelihood-free estimation method (Gutmann and Corander 2016; Jokinen et al. 2021; Kangasrääsiö et al. 2019). The likelihood is computed based on the model M of the user interacting with the task environment ϕ. The posterior of parameters can be used to assess what caused the observed behavior, such as subjective utility and capacities.

What if? questions: The agent attempts to optimize a set of interventions $i^* \subset \mathcal{D}$ (often called "designs"), that out of the space of all possible designs maximize the expected value function $\mathcal{V}$, given the predicted history of behavioral sequences that are adaptations to these interventions by the user:

$$i^* = \arg\max_{i} \mathcal{V}(h), \qquad (3)$$

where $h \sim M(\pi^*, \theta^*, \phi_i)$. Similarly as the utility function $\mathcal{U}$, the value function $\mathcal{V}$ maps a history h to a scalar. Determining $\mathcal{V}$ is problem-dependent, but for instance, one can set $\mathcal{V} = \mathcal{U}$, so that the AI has the same utility as the user.

WHAT IF? MAKING PREDICTIONS UNDER ADAPTATION

Computational rationality permits answering "what if?" questions by simulating the consequences of candidate interventions (Oulasvirta, Jokinen, and Howes 2022). Unique to the approach is that it allows taking the human adaptive response into account, including its costs, thereby allowing the AI to design optimal interventions.

Recently, applications have emerged in two areas: (1) computational design and (2) adaptive user interfaces (UIs). In computational design, computationally rational models have been used to determine design objectives (Equation 3 in Box 1). For example, models of sensorimotor

performance in typing have been used in the optimization of intelligent text entry support for users with dyslexia and tremor (Sarcar et al. 2018). Here, the design problem is how to organize letters on keys, how to size them, and how to offer word prediction support. Combinatorial optimization was used to find the best design by consulting a model that adapted its typing strategy—here speed–accuracy trade-off and proofreading frequency—to each candidate design; thereby answering "what if?" questions. Typing errors for users with tremor were predicted to reduce from 50% to around 5% on a keyboard where letter groups and word predictions were optimized. Approaches to computational design prior to computational rationality did not consider how users adapt their interaction as a function of their capabilities and goals, instead they assumed a fixed policy (e.g., Gajos, Wobbrock, and Weld 2008).

In applications of adaptive interfaces, the key challenge is to estimate the costs of *changing* the interface, given the user's history. A model of visual search and layout learning was used to individualize websites based on user history (Jokinen et al. 2020; Todi et al. 2019). The model simulated cognitive abilities, such as visual search and long-term memory using existing psychological models. Search policies were estimated that optimally exploited the available resources provided by the website design, given bounds such as foveated vision and limited memory. When user' histories with different layouts was considered, the impact of adjusting a layout could be calculated by simulating how the user adapts to the change. Websites optimized using this approach were 25% faster to search visually (Todi et al. 2019). In another example, computational rationality was used to explore how drivers might adapt their multitasking to different lane control implementations (Jokinen and Kujala 2021). The example highlights the importance of being able to model how humans adapt to changes in the task environment: increased automation, especially if the user trusts its capabilities, may result in drastic behavioral changes that need to be considered when adapting interaction.

Researchers have also looked at how to use "what if?" questions to help learn a model of the user's reward function (Reddy et al. 2020). Here, a user model synthesizes hypothetical behaviors, asks the human user to label the behaviors with rewards, and then trains a neural network to predict the rewards. The approach was tested on a navigation task and a video game. The results show that a method that asks "what if?" questions of a user model significantly outperforms prior methods in learning reward functions that transfer to new environments with different initial state distributions.

The ability to ask "what if?" questions is also a key element of approaches to automate the testing of psychological models. The idea is to estimate model parameters and choose between models by choosing optimal experiments for a human participant. Cavagnaro et al. (2013) demonstrated that querying a space of candidate models of human risky choice with proposed experimental stimuli could be used to generate optimal stimuli (mini-experiments) for discriminating between models. This approach attempts to choose the most diagnostic stimuli by estimating the consequences of potential stimuli using its model, in other words by conducting "what if?" experiments. By simulating many such hypothetical experiments, an expected utility of each stimulus was computed, and the stimulus with the highest expected utility was selected as the next to be given to the participant. While this approach makes no commitment to a particular utility function, the choice is critical to the success of its use. The particular form of the utility function used by Cavagnaro et al. (2013) was mutual information between a variable representing the uncertainty over the set of models under consideration and a variable representing the uncertainty associated with a particular miniexperiment (a single trial). This utility can be interpreted as the reduction in uncertainty that would be achieved by observing the outcome if the miniexperiment was actually conducted. However, no published examples of the use of this method have considered the utility to the experimental participant ($\mathcal{U}$ in Equation 1).

WHY? INFERENCE AND THE PROBLEM OF INVERSE MODELING

An ideal answer to a "why?" question might consist of a distribution over psychologically relevant latent constructs, including, (1) subjective utilities, (2) capacities, and (3) the adaptation environment. The problem for a cooperative agent is to calculate this distribution from observed behavior; a number of approaches to related problems are relevant.

One approach is inverse reinforcement learning (IRL) (Kangasrääsiö and Kaski 2018; Ng and Russell 2000); the reward function of an observed agent is inferred from observations of its behavior under the assumption that the behavior was generated with an optimal policy (Equation 2). A typical IRL algorithm takes as input a set of behavioral trajectories and outputs a reward function. The observed agent is modeled as an MDP and the reward function of this MDP is modeled as, for example, a distribution over possible rewards. Given an initialization, the MDP is solved for the current reward function to generate a policy and a behavior. IRL then updates the current reward function to minimize the divergence between the observed behavior and the behavior generated by the learned policy. This process is repeated until convergence,

or alternatively a posterior distribution is produced for estimating plausible reward functions (Jokinen et al. 2021).

IRL is claimed to be a route to learning from a human expert from demonstrations and could be applied to answering "why?" questions when the answer is in terms of human goals (Hadfield-Menell et al. 2016). However, there are a number of challenges (Arora and Doshi 2021; Zhifei and Meng Joo 2012). The computational costs of solving the problems tend to grow rapidly with the size of the problem. More importantly, from the perspective of the current paper, IRL is not developed to take into account latent factors other than the reward function; factors that are critical to modeling humans including cognitive limitations and experienced ecology, which are known causal determinants of behavior, that vary across people.

Another problem with IRL is that it assumes that the human policy is optimal with respect to the task at hand, which precludes the possibility that the human is optimal with respect to say, teaching the task; in other words, the human is performing the task in a way to make it easier for the IRL agent to learn. The human might position a hand in such a way so as to make an action visible, rather than so as to minimize movement time. This problem with IRL is addressed by cooperative inverse reinforcement learning (CIRL) (Hadfield-Menell et al. 2016), which as with IRL makes inferences about the reward function but does so by framing the problem as a cooperative partial-information game. CIRL, thereby, computes optimal joint policies.

A different approach to answering "why?" questions could be based on likelihood-free inference (LFI) methods. The problem of inferring a reward function, or a memory or attention parameter, is a problem of parameter inference (Kangasrääsiö et al. 2019). As computationally rational simulators of human cognition are not expressed as closed-form equations, they have no known closed-form likelihood functions. As a consequence, it is difficult to estimate the plausibility of parameter values. LFI has received increasing interest as a way to infer parameters of complex simulation models (Cranmer, Brehmer, and Louppe 2020; Turner and Van Zandt 2012; Lintusaari et al. 2017; Sisson, Fan, and Beaumont 2019; Kangasrääsiö et al. 2017; Kangasrääsiö et al. 2019).

In LFI, observations are compared to sampled predictions made using the simulator. This permits approximating the plausibility of parameter values without an explicit likelihood function. Although the likelihood-function is not known, it does exist, and drawing a large number of samples from the model can give sufficient information about the plausibility of different parameter values. For instance, approximate Bayesian computing (ABC) performs parameter inference by simulating a model multiple times with various parameter values, and rejecting all but the subset of simulated samples that are close to the observed data. Due to the high-dimensionality of data from most stochastic simulations, a practical step commonly used in LFI methods is to reduce dimensionality by creating summary statistics (Hartig et al. 2011). The choice of these summaries is a critical factor for successful inference: they must represent the original data *sufficiently*, that is, an ideal summary statistic is informationally equivalent to the original data, given the goal of inferring model parameters. Domain-specific knowledge is, therefore, generally necessary when reducing dimensionality of data, although there are some recent examples of doing this automatically (Chen et al. 2020). By running the simulator multiple times and computing the difference between the summary statistics from the observed data and generated data, one can arrive at an approximation of the likelihood of different parameter values.

The choice of summaries presents a critical challenge for machines that understand people (Jokinen et al. 2021; Kangasrääsiö et al. 2019; Kangasrääsiö and Kaski 2018). The high-dimensionality of observation data necessitates the summaries, but at the same time, the sufficiency of any given summary statistics is questionable exactly because they reduce dimensionality and therefore reduce informativeness of the data (Hartig et al. 2011; Turner and Sederberg 2014). The key is in finding metrics that serve cooperative objectives, and acknowledging that the summaries they provide are not sufficient in the strict mathematical sense, but that they are sufficient with regards to what is relevant for capturing the individual variation in task behavior and explaining it using parameters of cognitive models.

The approach of treating parameters of cognitive models as random quantities instead of fixed values has implications for the types of statistics that are best suited to perform LFI. Because answers to "why?" questions should be expressed in terms of how likely different causes of observed behavior are, parameter inference techniques that estimate a posterior are preferable to those that estimate a fixed point. In this regard, some techniques, such as ABC, seem especially well suited, because they use probability theory to express information and uncertainty about parameters (Turner and Van Zandt 2012). The outcome of such methods is a posterior distribution of plausible parameter values instead of a fixed point estimate, permitting machines to evaluate the probabilities of alternative responses to "why?" questions. The advantages are that they permit accounting of prior information about parameter values, exploring uncertainty about parameter estimates by directly expressing how plausible certain values are, and accumulating evidence throughout longer time periods (Myung and Pitt 2016). This is important in cases where incorrect answers might have drastic

outcomes, for instance, in terms of user adaptations that lead to unsafe interaction.

LFI was recently used to understand task interleaving (Gebhardt, Oulasvirta, and Hilliges 2020). It explains why some people are more proficient at making efficient task switching decisions by modeling task interleaving as a hierarchical RL problem, where each available task is associated with its own cost and reward estimates. Fitted to almost 200 individual participant data, the analysis suggested that just two parameters—the discount factor of the supervisory controller and a coefficient that controlled a trade-off between rewards and task switching costs— could explain much of individual differences observed; thereby answering a "why?" question. Users with high discount factor persisted longer in a task before switching to another task.

"Why?" questions were answered in work by Jokinen et al. (2021) in an effort to determine the causes of lane deviation in driving. The inference was used to determine the level of motor noise giving rise to observed deviations. Individually, inferred parameters were presented as a posterior over different values, indicating confidence in the inference, with the posteriors of some users being more concentrated than for others, thus expressing how much power the driving noise parameter had in explaining a particular driver's behavior. Moreover, these posteriors were then sampled to generate posterior predictions of driving, resulting in isolating individuals who were at high risk at producing "tail cases," that is, situations of extreme and therefore dangerous lane deviations. This ties back to asking "what if?" in terms of probabilities. Combining computationally rational models with LFI significantly enhances the power of both for explaining human behavior.

Answering "why?" questions is rarely a one-shot exercise with given data, rather data are gathered through a sequence of interventions, or experiments. A problem then concerns how to design an optimal sequence of experiments. This problem has been pursued in cognitive science (Cavagnaro et al. 2013; Myung and Pitt 2016; Turner, Sederberg, and McClelland 2016) and in statistics (Ryan et al. 2016). It sometimes referred to as "adaptive design optimization," sometimes as "optimal design." It provides methods for the allocation of experiments in an information gathering procedure. While there is interest in optimal experimental design across the sciences, there is particular interest in cognitive science. Optimal experimental designs may be used to achieve parameter estimation more efficiently and therefore answer "why?" questions more quickly and with lower cost to the human. Bayesian methods for optimal experimental design have been extensively explored (Ryan et al. 2016). One feature of Bayesian design methods is that the prior distribution can be updated after

only a single experiment (intervention) and observation. Bayesian experimental design requires the definition of a utility function, which represents the value of experiments. The Bayes optimal design maximizes the expected utility function over the design space. One of the most commonly used utility function is mutual information, which is used for both parameter estimation and model discrimination. However, mutual information is not straightforward to calculate and Bayesian design has mostly been limited to simple models because of the computational challenges of performing the maximization.

These issues have recently been addressed using amortized methods (Foster et al. 2019, 2021; Valentin et al. 2021; Ivanova et al. 2021). These involve learning a design policy upfront and then deploying this policy to make rapid design decisions at the time of experimentation. In other words, the computational challenge of Bayesian design is addressed by precomputing the design policy. A different approach uses a neural network to estimate a lower-bound of the mutual information—which is more cost-effective than the more constrained problem (Kleinegesse and Gutmann 2020; Gutmann and Corander 2016).

EXAMPLE: BUILDING COMPUTATIONALLY RATIONAL MODELS

In this section, we illustrate three key steps in a workflow for building computationally rational models, and exemplify it with a model:

1. Specify the external environment of the user.
2. Specify the internal environment as an MDP (Markovian decision process). It describes the user's cognitive processes and the reward function.
3. Estimate cognitive parameters from human data.

As the working example, we use a driver model created for intelligent lane assistance in semi-autonomous vehicles. In intelligent lane assistance, the ideal timing and degree of assistance depend both on the traffic situation and the driver's state. Assistive actions should be taken with consideration of the adaptive response of the driver given this state. Importantly, the driver's response depends on how much they trust the assistance. While this illustrative example is hypothetical, it builds on recent papers on the topic (Jokinen, Kujala, and Oulasvirta 2020; Jokinen and Kujala 2021; Jokinen et al. 2021).

Step 1.
The first step of the workflow is to formalize the external environment. This is a radical departure from data-driven

approaches, which usually begin with acquiring training data. The definition of the external environment involves modeling the states and transition dynamics of the external environment, and how they are affected by the actions of both the human user and the computer agent. The transition function can be modeled as hand-coded state transitions or via a physics simulator (e.g., Ikkala et al. 2022). In the driving model, the external environment (ϕ) would contain those aspects of the cockpit and traffic environment that are important for lane keeping, such as the road, signs, relative positions and speeds of vehicles, and the dynamics of the traffic flow. It may also contain steering controls and driving automation needed for lane assist. The level of detail in the model of the external environment depends on the goals of the modeling. It might be a high-level symbolic model and may, or may not, provide pixel-level renderings.

Step 2.

The second step is to specify the internal environment, which describes the user's cognitive processes and states. The internal environment is specified as an MDP. This step involves operationalizing psychological hypotheses about how the human mind processes task-relevant information. This requires answering questions such as how human senses provide information, how this information is integrated as goal-directed mental representations, and how those representations are stored, processed, and retrieved when making decisions. Previous work has modeled a number of cognitive, motor, and perceptual bounds (see Oulasvirta, Jokinen, and Howes (2022) for a review).

In the driver model, the internal environment simulates how a human driver perceives the external environment, and uses "percepts" to form an internal representation of the state. The representation integrates the perceptual information with prior representations of the task using an internal model of the task environment (Jokinen, Kujala, and Oulasvirta 2020). To account for more complex situations in driving, we may also need to model beliefs about the goals and behaviors of other drivers and a model of naive physics, or how drivers understand physics related to driving. Furthermore, beliefs such as how the driver represents the internal workings of the driving automation, are encoded here. Parameters θ are used to encode individual differences related to the internal environment; they can, for example, capture how novice versus expert drivers mentally represent and simulate the environment.

As part of this step, one needs to specify the reward function. While humans have complex, multifaceted preferences and life goals, the modeler should consider mainly those goals that are relevant to the task at hand.

Previous work has modeled a number of factors relevant for interactive tasks, such as efficiency (Chen 2015), efficacy (Jokinen, Kujala, and Oulasvirta 2020), ergonomics (Ikkala et al. 2022), or intrinsic motivation (e.g., Biehl et al. 2018; Singh et al. 2010). In complicated cooperative tasks, the reward function may include goals that conflict with those of the collaborator and which need to be managed or resolved.

In the driver model, the reward function quantifies the driver's aim to reach the intended destination safely and efficiently. In addition, a driver may have preferences about how they can best enjoy the trip, including music or the choice of alternative routes. These preferences will often be in conflict; for instance, the driver may need to consult an in-car interface for navigation or selection of the song, but this results in glances out of road and reduces safety. The resulting reward function should express how a driver might trade-off such objectives. Such preferences are encoded as weight parameters included in θ.

Specification of the internal and external environment also demands definition of the stimulus function that presents the external state to the internal environment, and the response function which transmits internal state to the external environment, largely through movement. Given the full model, the agent is trained using RL. It is trained through interaction with the model of the external environment. In the computational rationality workflow, unlike in earlier cognitive architecture models, the modeler does not need to predefine the policy. Instead, it is assumed that the agent's policy is optimal (π^*) within its bounds. The parameters θ are initially provided with a prior based on the modeler's assumptions about the values that researchers may have. At this stage in the workflow, the model is complete and initial steps can be taken to check that the policy converges to reasonable behaviors for sampled values of the parameters. Moreover, the optimal policy adapts to the design of the task, for example, how reliable the lane assist is (Jokinen and Kujala 2021), meaning that the predicted adapted behavior depends on the interventions that the AI chooses (elaborated below).

Step 3.

The third step in the workflow is parameter estimation. For instance, there may be a need to fit the model parameters θ to each individual human driver, or to a particular class of drivers. Both the internal and external environments have parameters that govern variation in the task and user behavior. Usually, the parameters of the external environment are well-known, but those of the internal environment are not. For example, the goals of the driver, how they process information and what knowledge they have, are unknown and must be inferred from their individual performance data. To facilitate

parameter estimation, the modeler can use a theoretical or empirical prior about plausible parameter values in the user population. Such priors can guide the use of the model during the early stages of interaction. However, as more observational data about a particular user are collected, cooperative AI must perform parameter inference in order to establish the most likely parameter estimates for that user. The main goal of this step is to define the parameters, how they are inferred from the data, and what priors to use, in order to accurately model the human user.

Parameter estimation can be extremely expensive computationally, especially when full Bayesian posteriors are required. Here, an additional step in the workflow can involve training the model, in silico, on a distribution of possible user parameters (Moon et al. 2022; Moon, Oulasvirta, and Lee 2023; Kwon et al. 2020) and even training the cooperative AI to acquire user data that is maximally useful for parameter estimation (Ryan et al. 2016; Foster et al. 2019; Ivanova et al. 2021; Keurulainen et al. 2023).

How could a model built with this workflow be tested and deployed? The first part of the answer to this question, is *very carefully* and responsibly, taking full account of the potential dangers, and ensuring oversight at every step (Yeung, Howes, and Pogrebna 2020). Technically, the goal of a cooperative AI system is to design an optimal set of interventions $i^* \subset \mathcal{D}$ to maximize the $\mathcal{V}$. Practically, this happens by simulating the parameterized model and assessing what sorts of interactions maximize this value. A vehicle using the driver model observes actions made by the driver and infers a posterior θ^* that best describes the behavior. If the driver seems unsure of the correct lane, or if there is a relatively large sway in the driving wheel, then the AI can infer that the driver is not very experienced, and adjust accordingly by designing an intervention ($i \in \mathcal{D}$) to make driving safer and more efficient. It could boost the lane assist to help the driver to keep the lane, or provide the driver with information about the best lane, given where the driver is going. By searching or planning possible changes to the environment, and using predictions made by the fitted model, the AI agent can factor in how the human user adapts to particular design interventions, allowing it to determine the best set of design interventions. This is a potentially powerful feature of computational rational models, as it makes it possible to study how a particular human user adapts to various design interventions without the modeler having to specify beforehand how the human user acts. While such interventions remain, for the moment, problems for studying in the laboratory, we are hopeful that they are on the critical path to real-world systems in the future.

PROGRESS AND CHALLENGES

Computational rationality is a model-based, first-principles approach to understanding humans. We believe that it has the potential, in the long-term, to go beyond data-driven methods—thanks to combining machine learning with psychological assumptions about the data-generating process, the human.

Critical to computational rationality is the assumption of rationality. Predictions about behavior are derived from a policy that is optimally adapted to utility and the bounds—possible causes of behavior. The assumption is simple but powerful: Human behavior can be explained and predicted by assuming that people do what is best, given what they want and are able to do. The identifiability problem is reduced, though not solved, by excluding nonoptimal policies from the space of possible explanations. The causal link, made available by the optimization assumption, makes it possible to use the model to answer the "what if?" and "why?" questions that are essential to the evaluation of interventions as well as to updating the model when new evidence is observed. Theories of bounds can build on results in cognitive science, providing a strong prior that would otherwise make the inverse problem intractable.

Many significant challenges remain if this vision is to be fully realized, however. One of the most significant concerns the nature of *human motivation*: What makes a person pursue some activities and not others? What, for example, makes one person give up in the face of adversity, while others persist? One part of the answer is, as we have argued, that explanations of behavior must be a joint function of reward function, resource limits, and environment. But another answer rests on a deeper understanding of the human reward function itself. In computational models of *intrinsic motivation*, such as *curiosity*, for example, states that are novel to the agent are hypothesized to be rewarding in their own right (Schmidhuber 1991). However, psychological theory suggests an even more nuanced picture. Related to this is work on the rational basis of motivation (Dubey and Griffiths 2020a, 2020b; Biehl et al. 2018; Schmidhuber 2010; Roohi et al. 2018; Ecoffet et al. 2021). A rational agent should be motivated to explore stimuli that maximally increase the usefulness of its knowledge (Dubey and Griffiths 2020a). Curiosity is studied as a mechanism by which humans approximate this rational behavior. Further, self-determination theory (Deci and Ryan 2000) can help understand how motivations evolve dynamically over time, in a complex interplay of beliefs, experiences, and basic organismic needs.

Another challenge concerns the <u>human</u> ability to *generalize* knowledge and skills. Without understanding this

ability, AI will grossly underestimate human abilities in novel encounters, such as when facing an intervention the AI has taken. So far, most models of computational rationality use model-free RL, which does not perform well compared to model-based alternatives when the environment changes and transfer of knowledge and skills is required (Moerland, Broekens, and Jonker 2020). Research in cognitive sciences suggests that planning, memory, and supervisory control are exploited to overcome this limitation, such as in models of human cognition based on model-based (Doll, Simon, and Daw 2012), episodic (Gershman and Daw 2017), and hierarchical RL (Botvinick and Weinstein 2014). In probabilistic program induction, motor programs are learned by inducing them from experience and combining them like computer programs can combine scripts (Lake, Salakhutdinov, and Tenenbaum 2015). Future work should look at how generalization is achieved by combining symbolic (e.g., concepts) and other capabilities.

Computational rationality may also appear to be at odds with human emotions. However, this perceived conflict arises only if one associates emotions with maladaptive and irrational behavior, and contrasts them with rational goal-directed adaptive behavior (Volz and Hertwig 2016). The challenge emotions pose to computational rationality is that of treating emotion as part of rational adaptation, requiring a goal-directed account of emotion (Moors and Fischer 2019). A possible path towards a computationally rational account of emotion would exploit the connection between reward-prediction errors (RPEs) used in RL algorithms and the phasic activity of midbrain dopamine neurons, which seem to track RPEs in the human brain (Gershman and Uchida 2019; Mikhael et al. 2022). This approach could investigate the possibility that the mechanisms of rational adaptation and emotion are inseparable in humans. A collaborative agent equipped with such a model could predict not only the emotional responses of humans, but also how emotions might impact behavior in both positive and negative ways.

Social interaction poses another significant challenge to computational rationality. While there is increasing interest in MDP-based formalisms of social interaction, such as social MDPs (Tejwani et al. 2022), for applications in cooperative settings with humans, there is a need to model humans as special type of social interactors. In our account, human behavior in collaborative settings is adaptive: in social interaction, people adapt their expression to the distribution of relationships present in the audience. For example, when using social media, people exhibit a range of strategies in order to attempt to prevent context collapse (Davis and Jurgenson 2014). The basis of such adaptation should be modeled as beliefs about others, including their capabilities, personalities, intent, and

so forth. A scientific basis for this was laid in the theory of social cognition in the late 1980s (Fiske and Taylor 1991). The topic has experienced a revival in AI research in the study of social robotics (Wykowska, Chaminade, and Cheng 2016) and theory of mind in artificial agents (Nguyen and Gonzalez 2021).

Finally, we believe that computational rationality is a candidate technology on the path to human–AI alignment (Russell 2019). Russell outlines three principles of alignment: the machine (1) should maximize human utility, (2) is uncertain about the human utility function, and (3) should decrease this uncertainty by observing human behavior. While Russell presented IRL as a possible method for accomplishing this vision, a strong prior is required to push forward. IRL and cooperative IRL (Hadfield-Menell et al. 2016) will prove more effective if complemented with viable models not only of human utilities but also of human bounds. Rather than assuming that humans can be modeled as rational agents with varying reward functions, as in IRL, computational rationality also embraces variation in internal bounds, including variation in observation functions, internal state transitions, and environments of adaptation. The resulting optimization problems are more challenging but must be faced if IRL is to produce useful solutions to alignment problems.

CONCLUSION

We believe that model-free learning is insufficient for cooperative AI. Computational rationality offers a principled, model-based, basis for algorithms to drive both inference and planning in cooperative agents. It puts psychological constructs, similar to those that underlie human cooperative abilities, at the center of such algorithms. This enables the theory to disentangle causal contributions that a person's goals, capabilities, and environment make to their behavior. It offers an exciting avenue for research on cooperative agents that better understand humans, plan interventions, and make available explanations that are human-understandable.

CONFLICT OF INTEREST STATEMENT
The authors declare that there is no conflict.

ORCID
Antti Oulasvirta https://orcid.org/0000-0002-2498-7837

REFERENCES
Acharya, A., X. Chen, C. W. Myers, R. L. Lewis, and A. Howes. 2017. "Human Visual Search as a Deep Reinforcement Learning Solution to a Pomdp." In *CogSci*.

Albrecht, S. V., and P. Stone. 2018. "Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems." *Artificial Intelligence* 258: 66–95.

Anderson, J. R. 1990. *The Adaptive Character of Thought*. Psychology Press.

Anderson, J. R. 1991. "Is Human Cognition Adaptive?" *Behavioral and brain sciences* 14(3): 471–485.

Anderson, J. R. 2014. *Rules of the Mind*. New York: Psychology Press.

Anderson, J. R., C. F. Boyle, and B. J. Reiser. 1985. "Intelligent Tutoring Systems." *Science* 228: 456–62.

Ariely, D. 2009. *Predictably Irrational, Revised and Expanded Edition: The Hidden Forces that Shape our Decisions*. New York: Harper.

Arora, S., and P. Doshi. 2021. "A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress." *Artificial Intelligence* 297: 103500.

Bahrami, B., K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, and C. D. Frith. 2010. "Optimally Interacting Minds." *Science* 329: 1081–5.

Baker, C. L., J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. 2017. "Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing." *Nature Human Behaviour* 1: 1–10.

Barto, A. G., S. Singh, and N. Chentanez. 2004. "Intrinsically Motivated Learning of Hierarchical Collections of Skills." In *Proceedings of the 3rd International Conference on Development and Learning*, 112–19. Piscataway, NJ.

Batmaz, Z., A. Yurekli, A. Bilge, and C. Kaleli. 2019. "A Review on Deep Learning for Recommender Systems: Challenges and Remedies." *Artificial Intelligence Review* 52: 1–37.

Bellman, R. 1966. "Dynamic Programming." *Science* 153: 34–7.

Bhui, R., L. Lai, and S. J. Gershman. 2021. "Resource-Rational Decision Making." *Current Opinion in Behavioral Sciences* 41: 15–21.

Biehl, M., C. Guckelsberger, C. Salge, S. C. Smith, and D. Polani. 2018. "Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop." *Frontiers in Neurorobotics* 12: 45.

Botvinick, M., and A. Weinstein. 2014. "Model-Based Hierarchical Reinforcement Learning and Human Action Control." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 20130480.

Bowers, J. S., and C. J. Davis. 2012. "Bayesian Just-So Stories in Psychology and Neuroscience." *Psychological Bulletin* 138: 389-414.

Bridgers, S., J. Jara-Ettinger, and H. Gweon. 2020. "Young Children Consider the Expected Utility of Others' Learning to Decide What to Teach." *Nature Human Behaviour* 4: 144–52.

Brumby, D. P., D. D. Salvucci, and A. Howes. 2007. "Dialing while Driving? A bounded Rational Analysis of Concurrent Multi-Task Behavior." In *Proceedings of the 8th International Conference on Cognitive Modeling*, 121–6.

Callaway, F., R. Antonio, and G. Tom. 2020. "Fixation Patterns in Simple Choice are Consistent with Optimal Use of Cognitive Resources." *PsyArXiv preprint PsyArXiv*:https://doi.org/10.31234/osf.io/57v6k.

Callaway, F., A. Rangel, and T. L. Griffiths. 2021. "Fixation Patterns in Simple Choice Reflect Optimal Information Sampling." *PLoS Computational Biology* 17(3): e1008863.

Callaway, F., B. van Opheusden, S. Gul, P. Das, P. M. Krueger, T. L. Griffiths, and F. Lieder. 2022. "Rational Use of Cognitive Resources in Human Planning." *Nature Human Behaviour* 6: 1–14.

Cavagnaro, D. R., R. Gonzalez, J. I. Myung, and M. A. Pitt. 2013. "Optimal Decision Stimuli for Risky Choice Experiments: An Adaptive Approach." *Management Science* 59: 358–75.

Chase, V. M., R. Hertwig, and G. Gigerenzer. 1998. "Visions of Rationality?" *Trends in Cognitive Sciences* 2: 206–14. http://www.ncbi.nlm.nih.gov/pubmed/21227174.

Chen, X. 2015. "An Optimal Control Approach to Testing Theories of Human Information Processing Constraints." PhD thesis, University of Birmingham.

Chen, X., A. Acharya, A. Oulasvirta, and A. Howes. 2021. "An Adaptive Model of Gaze-Based Selection." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11.

Chen, X., G. Bailly, D. P. Brumby, A. Oulasvirta, and A. Howes. 2015. "The Emergence of Interactive Behaviour: A Model of Rational Menu Search." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4217–26. ACM.

Chen, H., H. J. Chang, and A. Howes. 2021. "Apparently Irrational Choice as Optimal Sequential Decision Making." In *AAAI'21 Proceedings of the Thirty-fifth AAAI Conference on Artificial Intelligence*. AAAI Press.

Chen, X., and A. Howes. 2012. "A Reinforcement Learning Model of Bounded Optimal Strategy Learning." In *International Conference on Cognitive Modeling*. Berlin.

Chen, X., S. D. Starke, C. Baber, and A. Howes. 2017. "A Cognitive Model of How People Make Decisions Through Interaction With Visual Displays." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1205–16. ACM.

Chen, Y., D. Zhang, M. Gutmann, A. Courville, and Z. Zhu. 2020. "Neural Approximate Sufficient Statistics for Implicit Models." *arXiv preprint arXiv:2010.10079*.

Colombo, M., L. Elkin, and S. Hartmann. 2021. "Being Realist About Bayes, and the Predictive Processing Theory of Mind." *The British Journal for the Philosophy of Science*.

Consul, S., L. Heindrich, J. Stojcheski, and F. Lieder. 2022. "Improving Human Decision-Making by Discovering Efficient Strategies for Hierarchical Planning." *Computational Brain & Behavior* 5(2): 185–216.

Conway, A., C. Jarrold, and A. Miyake. 2008. *Variation in Working Memory*. New York: Oxford University Press.

Cowan, N. 2010. "The Magical Mystery Four: How is Working Memory Capacity Limited, and Why?" *Current Directions in Psychological Science* 19: 51–7.

Cranmer, K., J. Brehmer, and G. Louppe. 2020. "The Frontier of Simulation-Based Inference." *Proceedings of the National Academy of Sciences* 117: 30055–62.

Dafoe, A., Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. 2021. "Cooperative AI: Machines Must Learn to Find Common Ground." *Nature* 593(7857): 33–36.

Davis, J. L., and N. Jurgenson. 2014. "Context Collapse: Theorizing Context Collusions and Collisions." *Information, Communication & Society* 17: 476–85.

Daw, N. D. 2014. "Advanced Reinforcement Learning." *Neuroeconomics* 299–320.

Dayan, P., and N. D. Daw. 2008. "Decision Theory, Reinforcement Learning, and the Brain." *Cognitive, Affective & Behavioral Neuroscience* 8: 429–53. http://www.ncbi.nlm.nih.gov/pubmed/19033240.

Deci, E. L., and R. M. Ryan. 2000. "The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior." *Psychological Inquiry* 11: 227–68.

Doll, B. B., D. A. Simon, and N. D. Daw. 2012. "The Ubiquity of Model-Based Reinforcement Learning." *Current Opinion in Neurobiology* 22: 1075–81.

Dubey, R., and T. L. Griffiths. 2020a. "Reconciling Novelty and Complexity Through a Rational Analysis of Curiosity." *Psychological Review* 127: 455–76.

Dubey, R., and T. L. Griffiths. 2020b. "Understanding Exploration in Humans and Machines by Formalizing the Function of Curiosity." *Current Opinion in Behavioral Sciences* 35: 118–24.

Duggan, G. B., and S. J. Payne. 2009. "Text Skimming: The Process and Effectiveness of Foraging Through Text Under Time Pressure." *Journal of Experimental Psychology: Applied* 15: 228–42.

Ecoffet, A., J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. 2021. "First Return, then Explore." *Nature* 590: 580–6.

Fisac, J. F., M. A. Gates, J. B. Hamrick, C. Liu, D. Hadfield-Menell, M. Palaniappan, D. Malik, S. S. Sastry, T. L. Griffiths, and A. D. Dragan. 2020. "Pragmatic-Pedagogic Value Alignment." In *Robotics Research*, 49–57. Cham, Switzerland.

Fiske, S. T., and S. E. Taylor. 1991. *Social Cognition*. New York: Mcgraw-Hill Book Company.

Forbus, K. D., and T. Hinrich. 2017. "Analogy and Relational Representations in the Companion Cognitive Architecture." *AI Magazine* 38: 34–42.

Foster, A., D. R. Ivanova, I. Malik, and T. Rainforth. 2021. "Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design." In *International Conference on Machine Learning*, 3384–95. PMLR.

Foster, A., M. Jankowiak, E. Bingham, P. Horsfall, Y. W. Teh, T. Rainforth, and N. Goodman. 2019. "Variational Bayesian Optimal Experimental Design." *Advances in Neural Information Processing Systems* 32.

Gajos, K. Z., J. O. Wobbrock, and D. S. Weld. 2008. "Improving the Performance of Motor-Impaired Users with Automatically-Generated, Ability-Based Interfaces." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1257–66.

Gebhardt, C., A. Oulasvirta, and O. Hilliges. 2020. "Hierarchical Reinforcement Learning Explains Task Interleaving Behavior." *Computational Brain & Behavior* 4: 1–21.

S Geisler, W. 2011. "Contributions of Ideal Observer Theory to Vision Research." *Vision Research* 51: 771–81.

Gershman, S. J., and N. D. Daw. 2017. "Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework." *Annual Review of Psychology* 68: 101–28.

Gershman, S. J., E. J. Horvitz, and J. B. Tenenbaum. 2015. "Computational Rationality: A Converging Paradigm for Intelligence in Brains, Minds, and Machines." *Science* 349: 273–8.

Gershman, S. J., and N. Uchida. 2019. "Believing in Dopamine." *Nature Reviews Neuroscience* 20: 703–14.

Gigerenzer, G. 2018. "The Bias in Behavioral Economics." *Review of Behavioral Economics* 5: 303–36.

Gutmann, M. U., and J. Corander. 2016. "Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models." *Journal of Machine Learning Research* 17, 1–47.

Hadfield-Menell, D., A. Dragan, P. Abbeel, and S. Russell. 2016. "Cooperative Inverse Reinforcement Learning." *arXiv preprint arXiv:1606.03137*.

Hahn, U., and P. Warren. 2009. "Perceptions of Randomness: Why Three Heads are Better than Four." *Psychological Review* 116: 454–61.

Hartig, F., J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. 2011. "Statistical Inference for Stochastic Simulation Models—Theory and Application." *Ecology Letters* 14: 816–27.

Howes, A., R. L. Lewis, and A. Vera. 2009. "Rational Adaptation Under Task and Processing Constraints: Implications for Testing Theories of Cognition and Action." *Psychological Review* 116: 717–51.

Howes, A., P. Warren, G. Farmer, W. El-Deredy, R. Lewis. 2016. "Why Contextual Preference Reversals Maximize Expected Value." *Psychological Review* 123: 368–91.

Howes, A., P. A. Warren, G. Farmer, W. El-Deredy, and R. L. Lewis. 2016. "Why Contextual Preference Reversals Maximize Expected Value." *Psychological Review* 123: 368–91.

Ikkala, A., F. Fischer, M. Klar, M. Bachinski, A. Fleig, A. Howes, P. Hämäläinen, J. Müller, R. Murray-Smith, and A. Oulasvirta. 2022. "Breathing Life Into Biomechanical User Models." In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–14.

Ivanova, D. R., A. Foster, S. Kleinegesse, M. U. Gutmann, and T. Rainforth. 2021. "Implicit Deep Adaptive Design: Policy-Based Experimental Design Without Likelihoods." *Advances in Neural Information Processing Systems* 34: 25785–98.

Jara-Ettinger, J. 2019. "Theory of Mind as Inverse Reinforcement Learning." *Current Opinion in Behavioral Sciences* 29: 105–10.

Jara-Ettinger, J., L. E. Schulz, and J. B. Tenenbaum. 2020. "The Naive Utility Calculus as a Unified, Quantitative Framework for Action Understanding." *Cognitive Psychology* 123: 101334.

Johnson, S. G., and L. J. Rips. 2015. "Do the Right Thing: The Assumption of Optimality in Lay Decision Theory and Causal Judgment." *Cognitive Psychology* 77: 42–76.

Jokinen, J., A. Acharya, M. Uzair, X. Jiang, and A. Oulasvirta. 2021. "Touchscreen Typing as Optimal Supervisory Control." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Jokinen, J. P., and T. Kujala. 2021. "Modelling Drivers' Adaptation to Assistance Systems." In *Automotive UI*, 12–19.

Jokinen, J. P., T. Kujala, and A. Oulasvirta. 2020. "Multitasking in Driving as Optimal Adaptation Under Uncertainty." *Human Factors* 63(8).

Jokinen, J. P., U. Remes, T. Kujala, and J. Corander. 2021. "Bayesian Parameter Inference for Cognitive Simulators." In *Bayesian Methods for Interaction Design*, edited by N. Banovic, P. O. Kirstensson, A. Oulasvirta, and J. H. Williamson. Cambridge: Cambridge University Press.

Jokinen, J. P., Z. Wang, S. Sarcar, A. Oulasvirta, and X. Ren. 2020. "Adaptive Feature Guidance: Modelling Visual Search With Graphical Layouts." *International Journal of Human-Computer Studies* 136: 102376.

Juechems, K., J. Balaguer, B. Spitzer, and C. Summerfield. 2021. "Optimal Utility and Probability Functions for Agents with Finite Computational Precision." *Proceedings of the National Academy of Sciences* 118: e2002232118.

Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica: Journal of the Econometric Society* 47: 263–91.

Kangasrääsiö, Antti, Kumaripaba Athukorala, Andrew Howes, Jukka Corander, Samuel Kaski, Antti Oulasvirta. 2017. "Inverse Modeling of Complex Interactive Behavior with ABC." In *Proceedings of the ACM CHI'17 Conference on Human Factors in Computing Systems*.

Kangasrääsiö, A., and S. Kaski. 2018. "Inverse Reinforcement Learning from Summary Data." *Machine Learning* 107: 1517–35.

Kangasrääsiö, A., J. P. Jokinen, A. Oulasvirta, A. Howes, and S. Kaski. 2019. "Parameter Inference for Computational Cognitive Models with Approximate Bayesian Computation." *Cognitive Science* 43: e12738.

Keurulainen, A., I. Westerlund, O. Keurulainen, and A. Howes. 2023. "Amortised Experimental Design and Parameter Estimation for User Models of Pointing: A Reinforcement Learning Approach." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21.

Kleinegesse, S., and M. U. Gutmann. 2020. "Bayesian Experimental Design for Implicit Models by Mutual Information Neural Estimation." In *International Conference on Machine Learning*, 5316–26. PMLR.

Körding, K. P., and D. M. Wolpert. 2006. "Bayesian Decision Theory in Sensorimotor Control." *Trends in Cognitive Sciences* 10: 319–26.

Kwon, M., S. Daptardar, P. R. Schrater, and X. Pitkow. 2020. "Inverse Rational Control with Partially Observable Continuous Nonlinear Dynamics." *Advances in Neural Information Processing Systems* 33: 7898–909.

Laird, J. E. 2019. *The Soar Cognitive Architecture*. MIT Press, Cambridge, MA.

Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum. 2015. "Human-Level Concept Learning Through Probabilistic Program Induction." *Science* 350: 1332–8.

Lake, B. M., T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. 2017. "Building Machines that Learn and Think Like People." *Behavioral and Brain Sciences* 40: E253.

Lewis, R. L., A. Howes, and S. Singh. 2014. "Computational rationality: Linking mechanism and behavior through bounded utility maximization." *Topics in cognitive science* 6: 279–311.

Lieder, F., and T. L. Griffiths. 2019. "Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences* 4: 1–85.

Lieder, F., P. M. Krueger, and T. Griffiths. 2017. "An Automatic Method for Discovering Rational Heuristics for Risky Choice." In *CogSci*.

Lintusaari, J., M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. 2017. "Fundamentals and Recent Developments in Approximate Bayesian Computation." *Systematic Biology* 66: e66–82.

Littman, M. L. 2015. "Reinforcement Learning Improves Behaviour from Evaluative Feedback." *Nature* 521: 445–51.

Luan, S., J. Reb, and G. Gigerenzer. 2019. "Ecological Rationality: Fast-and-Frugal Heuristics for Managerial Decision Making Under Uncertainty." *Academy of Management Journal* 62: 1735–59.

Maloney, L. T. T., and P. Mamassian. 2009. "Bayesian Decision Theory as a Model of Human Visual Perception: Testing Bayesian Transfer." *Visual Neuroscience* 26: 147–55.

Mata, R., R. Frey, D. Richter, J. Schupp, and R. Hertwig. 2018. "Risk Preference: A View from Psychology." *Journal of Economic Perspectives* 32: 155–72.

McClelland, J., and M. Botvinick. "Deep Learning: Implications for Human Learning and Memory." psyarxiv. https://psyarxiv.com/3m5sb/download?format=pdf

Meyer, D., and D. Kieras. 1997. "A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance: Part 1. Basic Mechanisms." *Psychological Review* 104: 3–65.

Mikhael, J. G., H. R. Kim, N. Uchida, and S. J. Gershman. 2022. "The Role of State Uncertainty in the Dynamics of Dopamine." *Current Biology* 32: 1077–87.

Moerland, T. M., J. Broekens, and C. M. Jonker. 2023. "Model-Based Reinforcement Learning: A Survey." *Foundations and Trends® in Machine Learning* 16(1): 1–118.

Moon, H.-S., S. Do, W. Kim, J. Seo, M. Chang, and B. Lee. 2022. "Speeding Up Inference with User Simulators Through Policy Modulation." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–21.

Moon, H.-S., A. Oulasvirta, and B. Lee. 2023. "Amortized Inference with User Simulations." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20.

Moors, A., and M. Fischer. 2019. "Demystifying the Role of Emotion in Behaviour: Toward a Goal-Directed Account." *Cognition and Emotion* 33: 94–100.

Myung, J. I., and M. A. Pitt. 2016. "Model Comparison in Psychology." *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, volume 5, Wiley.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard Univ Press.

Ng, A. Y., and S. J. Russell. 2000. "Algorithms for Inverse Reinforcement Learning." In *International Conference on Machine Learning*, 663–70.

Nguyen, T. N., and C. Gonzalez. 2022. "Theory of Mind from Observation in Cognitive Models and Humans." *Topics in Cognitive Science* 14(4): 665–686.

Oaksford, M., and N. Chater. 2007. *Bayesian Rationality: The probabilistic Approach to Human Reasoning*. Oxford, England: Oxford University Press.

Oulasvirta, A., J. P. Jokinen, and A. Howes. 2022. "Computational Rationality as a Theory of Interaction." In *CHI Conference on Human Factors in Computing Systems*, 1–14.

Pearl, J. 2021. "Radical Empiricism and Machine Learning Research." *Journal of Causal Inference* 9: 78–82.

Quinn, P., and S. Zhai. 2016. "A Cost-Benefit Study of Text Entry Suggestion Interaction." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 83–8.

Rahnev, D., and R. N. Denison. 2018. "Suboptimality in Perceptual Decision Making." *Behavioral and Brain Sciences* 41: e223.

Rakow, T., and B. R. Newell. 2010. "Degrees of Uncertainty: An overview and Framework for Future Research on Experience-Based Choice." *Journal of Behavioral Decision Making* 23: 1–14.

Rao, R. P. 2010. "Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes." *Frontiers in Computational Neuroscience* 4: 146.

Reddy, S., A. Dragan, S. Levine, S. Legg, and J. Leike. 2020. "Learning Human Objectives by Evaluating Hypothetical Behavior." In *International Conference on Machine Learning*, 8020–9. PMLR.

Rieskamp, J., and P. E. Otto. 2006. "SSL: A Theory of How People Learn to Select Strategies." *Journal of Experimental Psychology: General* 135: 207–36. http://www.ncbi.nlm.nih.gov/pubmed/16719651.

Roohi, S., J. Takatalo, C. Guckelsberger, and P. Hämäläinen. 2018. "Review of Intrinsic Motivation in Simulation-Based Game Testing." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Penguin.

Russell, S. J., and D. Subramanian. 1994. "Provably Bounded-Optimal Agents." *Journal of Artificial Intelligence Research* 2: 575–609.

Ryan, E. G., C. C. Drovandi, J. M. McGree, and A. N. Pettitt. 2016. "A Review of Modern Computational Algorithms for Bayesian Optimal Design." *International Statistical Review* 84: 128–54.

Sarcar, S., J. P. P. Jokinen, A. Oulasvirta, Z. Wang, C. Silpasuwanchai, and X. Ren. 2018. "Ability-Based Optimization of Touchscreen Interactions." *IEEE Pervasive Computing* 17: 15–26.

Schmidhuber, J. 1991. "A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers." In *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, 222–7.

Schmidhuber, J. 2010. "Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)." *IEEE Transactions on Autonomous Mental Development* 2: 230–47.

Simon, H.A. 1969. *The Sciences of the Artificial*. Cambridge: The MIT Press.

Singh, S., R. L. Lewis, and A. G. Barto. 2009. "Where Do Rewards Come From." In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2601–6. Cognitive Science Society.

Singh, S., R. Lewis, A. Barto, and J. Sorg. 2010. "Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective." *IEEE Transactions on Autonomous Mental Development* 2: 70–82.

Sisson, S. A., Y. Fan, and M. A. Beaumont, eds. 2019. *Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL.

Spektor, M. S., S. Gluth, L. Fontanesi, and J. Rieskamp. 2019. "How Similarity Between Choice Options Affects Decisions from Experience: The Accentuation-of-Differences Model." *Psychological Review* 126: 52–88.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press.

Swets, J., W. Tanner Jr, and T. Birdsall. 1961. "Decision Processes in Perception." *Psychological Review* 68: 301–40.

Tejwani, R., Y.-L. Kuo, T. Shu, B. Katz, and A. Barbu. 2022. "Social Interactions as Recursive MDPs." In *Conference on Robot Learning*, 949–58. PMLR.

Todi, K., J. Jokinen, K. Luyten, and A. Oulasvirta. 2019. "Individualising Graphical Layouts with Predictive Visual Search Models." *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10: 1–24.

Tseng, Y.-C., and A. Howes. 2015. "The Adaptation of Visual Search to Utility, Ecology and Design." *International Journal of Human-Computer Studies* 80: 45–55.

Turner, B. M., and P. B. Sederberg. 2014. "A Generalized, Likelihood-Free Method for Posterior Estimation." *Psychonomic Bulletin & Review* 21: 227–50.

Turner, B. M., P. B. Sederberg, and J. L. McClelland. 2016. "Bayesian Analysis of Simulation-Based Models." *Journal of Mathematical Psychology* 72: 191–9.

Turner, B. M., and T. Van Zandt. 2012. "A Tutorial on Approximate Bayesian Computation." *Journal of Mathematical Psychology* 56: 69–85.

Valentin, S., S. Kleinegesse, N. R. Bramley, M. U. Gutmann, and C. G. Lucas. 2021. "Bayesian Optimal Experimental Design for Simulator Models of Cognition." *arXiv preprint arXiv:2110.15632*.

Volz, K. G., and R. Hertwig. 2016. "Emotions and Decisions: Beyond Conceptual Vagueness and the Rationality Muddle." *Perspectives on Psychological Science* 11: 101–16.

Welford, A. T. 1952. "The Psychological Refractory Period and the Timing of High-Speed Performance—A Review and a Theory." *British Journal of Psychology* 43: 2–19.

Wykowska, A., T. Chaminade, and G. Cheng. 2016. "Embodied Artificial Agents for Understanding Human Social Cognition." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371: 20150375.

Yeung, K., A. Howes, and G. Pogrebna. 2020. "AI Governance by Human Rights–Centered Design, Deliberation, and Oversight." *The Oxford Handbook of Ethics of AI*, 77–106. Oxford: Oxford Academic.

Zhifei, S., and E. Meng Joo. 2012. "A Survey of Inverse Reinforcement Learning Techniques." *International Journal of Intelligent Computing and Cybernetics* 5: 293–311.

AUTHOR BIOGRAPHIES

Andrew Howes is a Professor in the Department of Computer Science at University Exeter, UK. His research interest is in computational models of human decision making and rational adaptation. He has previously held academic posts at institutions that include the Medical Research Council, Cambridge, Carnegie Mellon University, Cardiff University, University of Manchester, University of Michigan, Aalto University, and University of Birmingham.

Jussi P.P. Jokinen is an Assistant Professor of Cognitive Science at Faculty of Information Technology, University of Jyväskylä, Finland. His research interest is in computational cognitive modeling of complex interactive tasks.

Antti Oulasvirta is a Professor in the Department of Information and Communications Engineering at Aalto University, where he leads the Computational Behavior Lab. He is a cognitive scientist with interest in modeling human behavior. Prior to joining Aalto, he was a Senior Researcher at the Max Planck Institute for Informatics. He serves as an Associate Editor for ACM TOCHI.