Cognitive science as a source of forward and inverse models of human decisions for robotics and control

Mark K. Ho[1] and Thomas L. Griffiths[1,2]

[1]*Princeton University, Department of Computer Science, Princeton, NJ, USA*

[2]*Princeton University, Department of Psychology, Princeton, NJ, USA*

Abstract

Those designing autonomous systems that interact with humans will invariably face questions about how humans think and make decisions. Fortunately, computational cognitive science offers insight into human decision-making using tools that will be familiar to those with backgrounds in optimization and control (e.g., probability theory, statistical machine learning, and reinforcement learning). Here, we review some of this work, focusing on how cognitive science can provide forward models of human decision-making and inverse models of how humans think about others' decision-making. We highlight relevant recent developments, including approaches that synthesize blackbox and theory-driven modeling, accounts that recast heuristics and biases as forms of bounded optimality, and models that characterize human theory of mind and communication in decision-theoretic terms. In doing so, we aim to provide readers with a glimpse of the range of frameworks, methodologies, and actionable insights that lie at the intersection of cognitive science and control research.

Keywords: cognitive science, robotics, psychology, decision-making, resource rationality, theory of mind

Contents

1 Introduction

As robots and other automated systems are beginning to become more integrated into human lives, engineers face a new problem: designing these systems to effectively and safely interact with people. Part of the challenge is that humans are themselves autonomous agents, making decisions and acting in ways that introduce potentially unpredictable dynamics into the environment. Even more challenging, humans change their behavior in response to the actions of the system they are interacting with, meaning that the engineer has to consider not just how to predict and interpret human behavior, but how the behavior of the system that they are designing might be predicted and interpreted by humans in turn.

As cognitive scientists, we have had many enjoyable conversations with engineers about how to solve these problems. Typically these conversations begin with an email or a knock on the door requesting the most up-to-date model of human behavior in a format that can be

easily integrated into a control-theoretic framework. We disappoint our colleagues by telling them that unfortunately no such model exists, but then excite them with how much progress has been made towards this goal and all of the research possibilities that this entails. Our intent in this article is to offer our readers a chance to follow the same emotional trajectory, highlighting the ways in which we think contemporary cognitive science can provide tools that may be useful to engineers designing systems that interact with humans and identifying some of the exciting possibilities for future research in this area.

Speaking broadly, computational models developed by cognitive scientists offer solutions to at least two of the problems that engineers face (Figure 1). First, they provide *forward models* that can be used to generate predictions about human behavior based on assumptions about the beliefs, goals, and desires of human agents. These models can be useful for anticipating what a person will do in a given situation and hence provide a way to enrich modeling of the environment in which an automated system operates. These forward models can also be used as an ingredient in *inverse models*, which infer the beliefs, goals, and desires of humans based on their actions – a necessary step for any agents that seek to coordinate their behavior with humans, collaborate effectively, provide assistance, or cooperate as they pursue common goals.

In addition to this, cognitive science also offers insight into the way that humans solve exactly this inverse problem. People routinely make inferences about the beliefs, goals, and desires of other people, a process that has been extensively studied by psychologists [1, 2, 3] and is increasingly captured in computational models [4, 5, 6]. These models are useful both as a source of insight for engineers seeking to re-create this capacity to draw inferences from the actions of others, but also as a tool for anticipating how the actions of an automated system will be interpreted by a human [7, 8]. Research in human-robot interaction has begun to make use of these ideas, designing systems that act in a way that is more legible to humans [9, 10, 11], a line of work that we will also review.

In considering these two ways that computational models of cognition can be used by

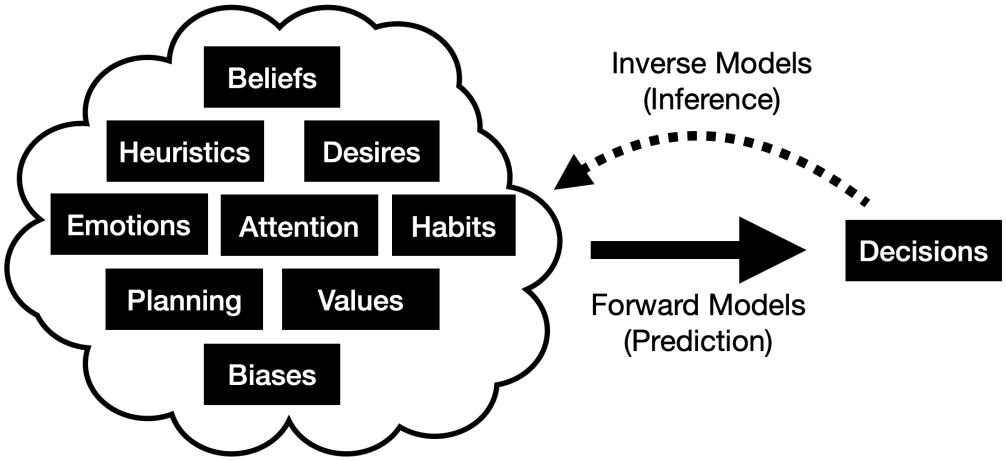

Figure 1: Forward and inverse models from cognitive science. In this paper, we review work in cognitive science on forward models of how humans make decisions and inverse models that humans use to reason about other agents. Cognitive scientists are increasingly using computational tools such as probability theory, reinforcement learning, and statistical machine learning to characterize forward and inverse models of human decision-making. This provides opportunities for cross-talk and collaboration between cognitive science and control research.

engineers designing automated systems – as both forward and inverse models – we will also highlight the ways in which recent work in computational cognitive science has emphasized formalisms that will be very familiar to researchers coming from a background of optimization and control. Cognitive scientists increasingly use ideas from probability theory, statistical machine learning, and reinforcement learning in specifying models of human cognition [12]. This creates an opportunity to develop a common language for describing the behavior of both humans and machines, and supports easier integration of insights from cognitive science into control.

Given the vast scope of human behavior, is necessary to limit our review to a specific subdomain of human activity. To that end, we will focus on models of human decision-making, broadly construed. The decisions people make reveal their preferences and determine their actions, key to the design of interactive systems. They also provide a rich territory for researchers, with formal models of human decision-making going back almost 300 years [13]. Our goal is to summarize the current state of the art in predicting and interpreting human decisions in a form that is immediately actionable by designers of automated systems.

The remainder of the paper is split into two parts. In the first part we focus on forward models, considering the criteria for useful computational models of human decision-making and summarizing recent research that aims to satisfy these criteria. We then turn to inverse models, describing the problem of inverse inference, summarizing the key ideas from the psychological literature, explaining how this has been translated into formal models, and highlighting some of the ways in which these ideas have been applied within robotics. We close with a brief discussion of some of the remaining open questions in these areas and possibilities for future research.

2 Forward models of human decision-making

For a theory of human decision-making to be useful to an engineer designing a system that has human behavior as a component, that theory should have two properties: it should be *generalizable*, meaning that it can be applied in any context in which the engineer needs to be able to make predictions about how people will act, and it should be *accurate*, producing good predictions about human behavior in that context. The development of theories of human decision-making has historically tended to alternate between these criteria, making progress on one at the cost of the other (Figure 2).

The earliest formal theories of human decision-making made the strong assumption that humans are rational, in the sense of pursuing actions that are in their self-interest and in compliance with axioms that can be widely agreed to characterize rational behavior [14, 15]. The impressive result of this investigation is that the preferences of a rational agent can be characterized by a utility function that assigns a numerical value to each possible outcome, and that when faced with decisions that involve uncertainty that agent should pursue the option that has highest expected utility.

This theory – which we will refer to as expected utility theory – fulfills the goal of generalizability. In order to predict the actions that a rational agent will take in a new

environment, it is necessary only to identify the utility assigned to different outcomes – the decisions that agent will take can then be derived directly from these quantities. The tools that are used for deriving this behavior are exactly the tools that are used in optimization and control, as we typically seek to define agents that are rational. As a consequence, human rationality is a common assumption in interactive systems, albeit with some allowance for stochasticity (e.g., [16]). It is also a common assumption in the kind of choice models that are widely used in econometrics (e.g., [17]).

The only problem with assuming that humans are rational is that this assumption turns out to be false. Starting in the 1970s, psychologists (led by Daniel Kahneman and Amos Tversky) began to document the ways in which people's decisions violate the axioms that are assumed in rational models [18]. This led to a swing towards a more qualitative psychology of human decision-making, in which the emphasis was placed on identifying all of the heuristic shortcuts that people seem to use when making decisions, and the behavioral biases that result. The outcome of this process is a long list of the things that people do wrong in specific situations. This is something that might potentially increase the accuracy of our models, but since these behaviors are specific to particular scenarios and it is hard to know which heuristic might dominate in a new setting, this accuracy comes at the cost of generalizability.

This qualitative approach to understanding human decision-making was complemented by efforts to formalize the cognitive processes that people engage in when making decisions. Kahneman and Tversky developed *prospect theory*, which extends expected utility theory by allowing different functions characterizing the subjective value of gains and losses and recognizing that the probabilities of events may also be subjectively transformed [23]. Subsequent work has introduced further nuances to this theory (e.g., [24]), together with hypotheses for how to formalize ideas about different heuristics that people might follow (e.g., [25]) as well as other cognitive factors such as the salience of different options (e.g., [26]).

Recent work in psychology and neuroscience has drilled down even further into the cognitive processes that might account for these behaviors. One prominent line of work focuses

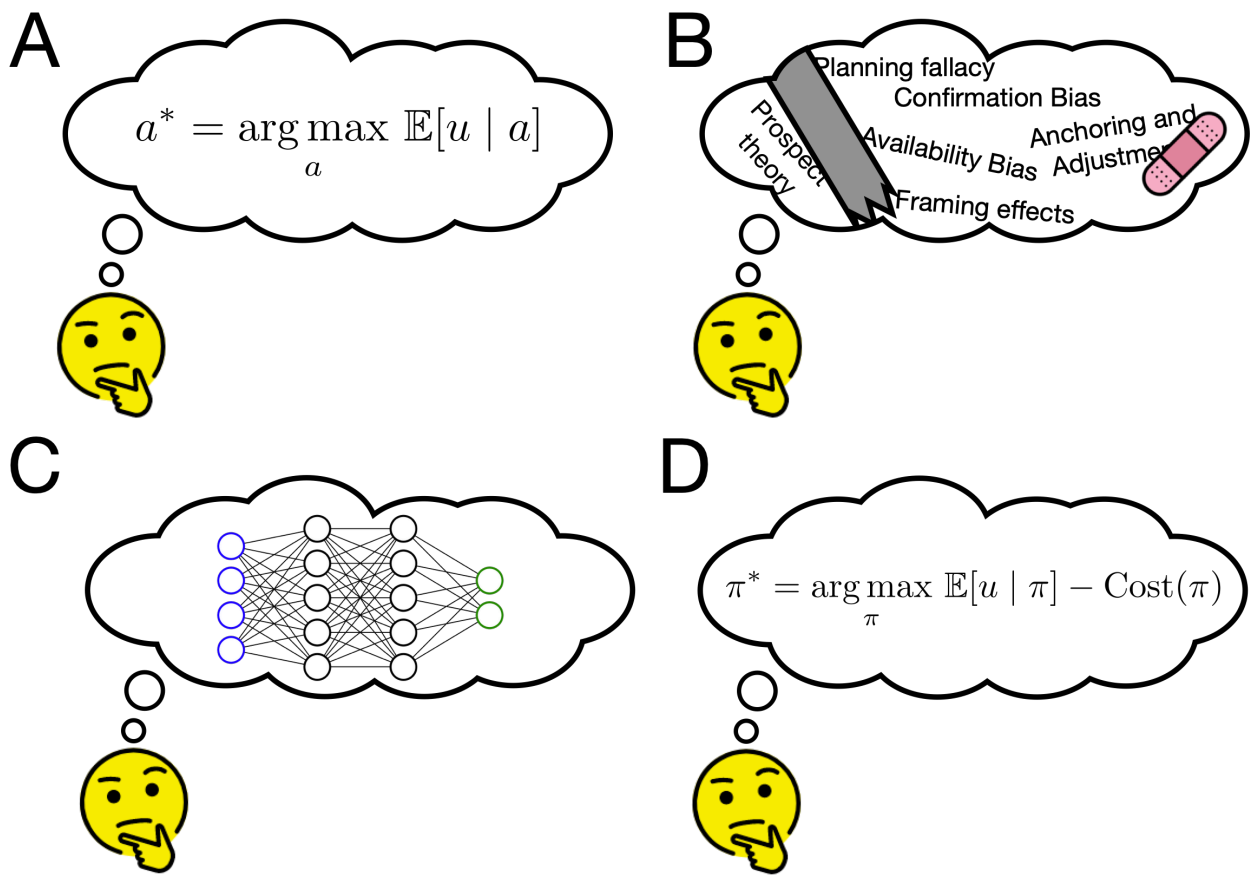

Figure 2: **Four approaches to studying human decision-making** (A) Early formal theories of decision-making assumed humans were *expected utility maximizers*. (B) The *heuristics and biases* research program initiated in the 1970's by Tversky and Kahneman [18] demonstrated that people systematically violate basic predictions of expected utility theory. This resulted in a focus on the heuristics that people tend to use in specific situations as opposed to general theories. (C) Decision-making models represented by neural networks, combined with large data sets of human choice behavior and informed by psychological theory (e.g., [19, 20]), provide a way to predict about human decisions (see section 2.1). (D) The theoretical framework of *resource rationality* [21, 22] aims to provide a general formal theory that accounts for people's heuristics and biases. Specifically, resource rationality proposes that human decision-making reflects expected utility maximization subject to computational costs and cognitive limitations (see section 2.2).

on the idea that people make decisions by accumulating evidence that one option is better than the alternatives [27]. There is ongoing debate about the precise mechanisms by which such a process could operate (e.g., [28]), but these accumulator models have also received support from results in neuroscience that seemed to show areas in the brain that engage in evidence accumulation (e.g., [29], but see [30]).

For the engineer, the precise cognitive and neural mechanisms underlying people's decisions might matter less than what those decisions actually are and how good predictions about human decision-making can be generated in other contexts – our two criteria of accuracy and generalizability. To this end, we are going to focus on two recent developments that increase the potential for models of human decision-making to be used effectively as forward models in control settings. The first is the potential to use ideas from machine learning, combined with the availability of large data sets on human behavior, to develop more accurate models of human decision-making. The second is recent efforts to revisit the notion of rationality, with the goal of obtaining a theoretical framework that has the same generalizability as expected utility theory while incorporating what we know about human cognitive limitations in a way that supports greater accuracy.

2.1 Developing more accurate models of human decisions using machine learning

The recent success of machine learning in many domains raises the possibility that such systems may be able to better predict human decisions than the theories of choice developed by psychologists and economists. For the engineer seeking a forward model of human decision-making, it may be tempting to collect a data set of human decisions and train an off-the-shelf machine learning method to predict people's behavior. This possibility has been explored extensively over the last decade, showing that machine learning has a great deal of promise in this area, but also that performance of these systems can be significantly improved by the injection of some psychological insight.

A first extensive comparison of psychological models against machine learning systems for predicting human decisions occurred in the 2015 Choice Prediction Competition [31]. The competition employed a standard *risky choice* paradigm that has been used extensively to study human decisions, informing the development of many of the models summarized above. In this task, participants make a choice between two gambles. In each gamble different outcomes – here corresponding to actual monetary gains and losses – occur with different probabilities. The pairs of gambles can be described by an 11 dimensional vector that summarizes the payoffs and their probabilities. The task is to map this 11 dimensional vector to a probability of choosing one gamble over the other, with the goal of getting this probability as close as possible to the choice probabilities of a group of human participants. In the competition, the choice probabilities for 90 such pairs of gambles were provided, and the goal was to predict the corresponding probabilities for a held-out test set.

The results of the 2015 Choice Prediction Competition showed that psychological models – that is, those developed by psychologists and economists – tended to outperform off-the-shelf machine learning methods. The best-performing model instantiated a set of heuristics that had been identified in the psychological literature. Subsequent work showed that machine learning methods could improve on this performance, but only when provided with features that were motivated by psychological theory [32, 33].

To machine learning practitioners, these results may not come as a big surprise. This prediction problem has a relatively large number of features compared to the amount of available data (90 pairs of gambles). The success of psychological models can be interpreted as an instance of the bias-variance trade-off [34], with the small amounts of data involved meaning that models with carefully crafted inductive biases are most likely be successful. However, the other side of that trade-off is the expectation that as the amount of data increases we should expect to see improved performance from machine learning models with weaker inductive biases.

Consistent with this hypothesis, applications of machine learning to predicting other kinds

of human decisions have shown greater success. With more instances of more constrained problems, machine learning methods can outperform psychological models, and have even been suggested as offering an upper bound on the amount of variance we can expect to account for [35, 36]. Indeed, in a subsequent Choice Production Competition where models were trained on 210 pairs of gambles, relatively generic machine learning models were able to outperform psychological theories [37].

Based on this insight, recent work has collected and analyzed a risky choice dataset that involves orders of magnitude more problems than the original Choice Prediction Competition [38, 19]. In this data set, human participants made decisions for over 10,000 pairs of gambles. The size of the data set makes it possible to systematically evaluate existing models of choice, and to use machine learning to exhaustively explore the space of possible theories. Different models of choice can be expressed in terms of constraints on the functional form of a predictive model. For example, under the expected etility theory we can take the probability that people choose a gamble to be proportional to $\exp\{\sum_i p_i u(x_i)\}$ where p_i is the probability of the outcome x_i and $u(\cdot)$ is a utility function. By taking an arbitrary differentiable form for this utility function – such as an artificial neural network – we can employ standard tools for automatic differentiation to use gradient descent to optimize the form of this function against human data. This approach generalizes to other psychological theories. For example, prospect theory corresponds to assuming the choice probability is proportional to $\exp\{\sum_i \pi(p_i) u(x_i)\}$ where $\pi(\cdot)$ is a probability weighting function.

Peterson et al. [19] used this approach to identify the optimal functional form for various classic theories of choice, and also evaluated unconstrained artificial neural networks for predicting people's decisions (see Figure 3). The results showed that when using the entire data set of different pairs of gambles, an unconstrained neural network systematically outperformed all existing psychological theories. However, they also showed that equivalent performance could be obtained by defining a model based on a mixture of these classic theories, and that this model achieved a high level of predictive performance far faster than an

unconstrained neural network. The results of this analysis suggest that, given enough data, we can obtain better forward models of human decisions using machine learning, but that these models are likely to be enhanced by drawing on psychological theory when possible.

A further challenge of using off-the-shelf machine learning methods when developing forward models of human decision-making is that these methods often result in models that are uninterpretable. Previous work in this area has relied on post hoc analysis of models to identify features that are psychologically interpretable (e.g., [36]). An alternative approach was recently outlined by Agrawal et al. [20], in which an off-the-shelf machine learning model is used to critique a more interpretable model until that interpretable model yields similar performance. This approach was applied to a large data set of human decisions that is likely to be of interest to researchers working on autonomous systems: the Moral Machine project [39]. This data set consists of more than 10 million human decisions about what an autonomous vehicle should do when faced with an inevitable collision, where the only available choice is about which group of pedestrians the vehicle will collide with (a version of the classic trolley problem [40]). Having an uninterpretable model that predicts these choices is not particularly useful for designing autonomous systems, but Agrawal et al. showed that their approach can be used to identify the features that a predictive model had discovered, making those features explicit in a way that is likely to be useful when deciding how to design and regulate autonomous vehicles.

2.2 Developing more generalizable models via resource rationality

Despite the promise of machine learning to improve the accuracy of a models of human decisions, generalizability is still likely to be a challenge. Machine learning systems are typically trained in a specific domain, and can face difficulty when applied in another related domain. Making this kind of generalization requires extracting the causal principles that underlie people's decisions. In this section we consider an approach that has the potential to do just that.

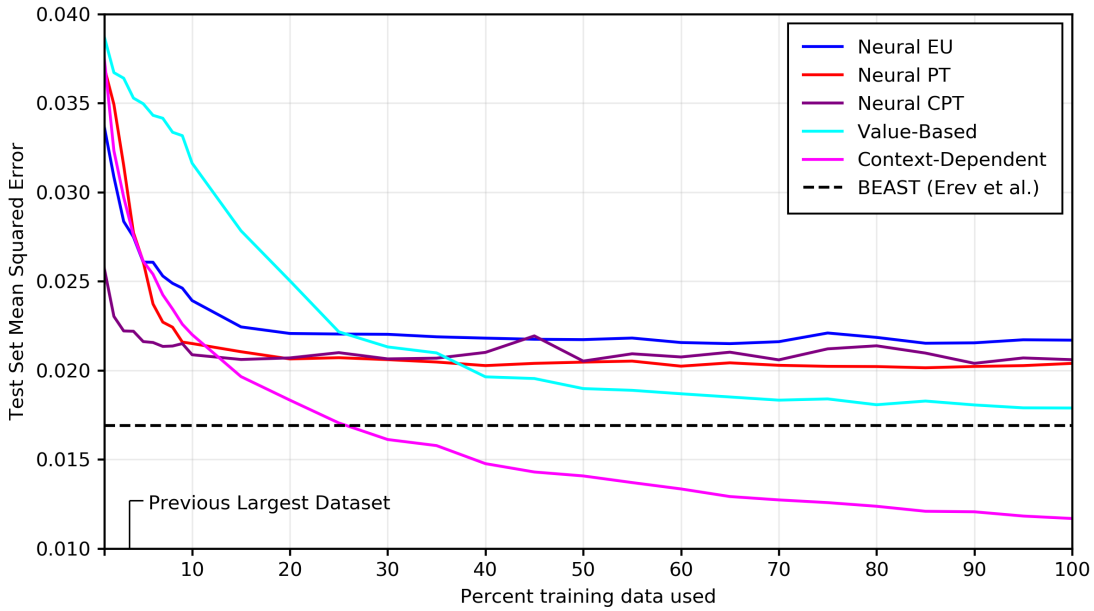

Figure 3: Performance of machine learning models protecting people's decisions in a risky choice task (data from [19]). Neural networks constrained to a functional form consistent with classic theories of decision-making such as expected utility (EU), prospect theory (PT), and cumulative prospect theory (CPT) are compared against networks that directly estimate the value of a gamble from its features (Value-Based) or directly predict people's choices based on the features of both gambles (Context- Dependent). The vertical axis shows means squared error in predicting the probability with which people choose a particular gamble, the horizontal axis shows the percent of The training data (approximately 10,000 pairs of gambles) that was used. The previous largest data set for risky choice [37] is shown. Given enough data, all neural network models outperform the best models in their class proposed by human psychologists and economists, but this requires orders of magnitude more data than have previously been collected. For comparison, the dotted line shows the performance of the Best Estimate And Sampling Tools (BEAST) model[31] that won the 2015 Choice Prediction Competition.

The classical notion of rationality was our prime example of a theory that satisfies the goal of generalizability – for any new situation, it is possible to derive predictions about behavior. However, this classical theory falls short not just because it fails to empirically capture aspects of how people make decisions, but because it represents an unrealistic ideal for any intelligent system with finite computational resources. This is a long-standing idea, going all the way back to the classic work of Herbert Simon on *bounded rationality* [41]. However, this idea has recently begun to receive a more comprehensive mathematical and empirical treatment.

The classical theory of rational action via maximizing expected utility doesn't take into account the computational cost of selecting that action. As a consequence, it's easy to imagine an agent trying to follow the prescriptions of this theory ending up paralyzed as it tries to compute all of the possible outcomes and their probabilities. To address this, researchers in the artificial intelligence literature have sought a more realistic criterion for rational action for agents with finite computational resources. The outcome of this investigation is the theory of bounded optimality, which focuses not on the optimal action that an agent should take but rather on the optimal algorithm an agent should follow in order to select that action [42, 43]. This theory explicitly trades off the expected utility of finally taking an action with the computational cost that's involved in getting to that point.

For cognitive scientists, bounded optimality offers a way to theorize about the optimal cognitive processes that intelligent agent should engage in when trying to make a decision [44, 45]. As it puts an emphasis on rational use of the cognitive resources an agent is able to apply, this approach has been referred to as *resource rational analysis* [21, 22]. Considering how an agent should rationally deploy its cognitive resources provides a way to explain why people may choose to adopt particular heuristics – even if those heuristics result in systematic biases – and to make generalizable predictions about the kinds of cognitive strategies that we expect people to engage in.

Recent research has recast some of the classic heuristics discovered by psychologists from

the perspective of the rational use of cognitive resources. For example, focusing on extreme events when considering the outcomes of a decision is a strategy that can minimize the variance of estimates of expected utility based on small samples, even though it introduces a bias to those estimates [46]. Thinking in these terms allows us to potentially begin to reconcile the various heuristics and biases identified by psychologists into a broader mathematical theory.

To the engineer, resource rationality offers the potential to make better predictions about human behavior in a way that incorporates realistic assumptions about the cognitive limitations of human agents. Generalizability results from the fact that deriving an optimal resource-rational strategy can be formulated as a sequential decision problem. An agent trying to make a decision is going to execute the sequence of computations that provide information about the possible outcomes of their actions, at some point selecting an action to perform based on this information. The sequential decision problem here corresponds to the choice of that sequence of computations: we can construe each computation as a kind of mental action, ending with the decision that we have done enough computation and are ready to act as the end of the sequence.

Expressed in these terms, it is possible to see that we can use familiar tools such as Markov decision processes to formalize the internal decision-making we do about how to deploy our cognitive resources. Solving the resulting MDPs (referred to as *meta-level MDPs* [47]) provides a way to derive predictions about behavior. Crucially, this approach carries with it the same generality as the classic theory of rational action. It simply moves rationality up the level of the choice of how to deploy cognitive resources, and requires us to be explicit about what those resources might be.

To provide a concrete example, one recent paper [48] used this approach to examine how can model attention allocation in a simple decision-making task. In this task people are presented with three objects – in this case snack foods – and asked to decide which they prefer. While they are doing this, their gaze it recorded using an eye tracker. People show

a consistent pattern of behavior in this task. For example, they spend more time looking at items that they assign a higher subjective value. These patterns of behavior can be explained by assuming that people are trying to estimate the subjective value of each item, and that each moment they spend looking at an option provides a sample from a Gaussian distribution centered on that value. The problem of deciding whether to sample can then be formulated as an MDP, and the resulting policy generates predictions about which objects people look at and in what sequence. We spend more time looking at items with higher subjective value because those are the items that are most relevant to our ultimate decision.

Describing cognitive processes in terms of the solution to Markov decision processes has the virtue of characterizing human cognition using formal tools that are likely to be familiar to those working in control theory. Other work on resource rationality has likewise employed formalisms that will be familiar to engineers. For example, one line of work focuses on the information-theoretic costs of maintaining mental representations at a given degree of precision [49, 50]. This approach also has connections with work in economics that explains apparently irrational aspects of human choice in terms of *rational inattention*, where information-theoretic costs are assumed to apply to the precision of the signal an agent uses to inform a decision [51, 52].

Resource rationality offers a new set of tools for capturing human behavior with greater precision in a way that is compatible with standard modeling techniques used in robotics and control. Being able to make generalizable predictions about how long it will take people to make a decision, what information they going to seek when making that decision, and what kinds of information are likely not to include when making a decision are all things that can facilitate the design of human-machine interfaces. Furthermore, it is possible to use this kind of approach to engage with the question of how to improve human decision-making: if we assume that people are rational but resource-limited, we can think about how assistive robots might modify the environments in which humans are making decisions to allow them to make better use of those resources.

3 Inverse models of human decision-making

While forward models can help us to make predictions about what people will do given their preferences, this is typically not all we need to know in order to design systems that are able to interact effectively with humans. In an ideal world, autonomous systems would effectively help fulfill people's needs and desires, which can change from person to person and situation to situation. We need a way to infer those needs and desires from people's behavior – inverse models. Furthermore, people are adaptive; they often change their behavior in response to a system based on their best guess as to how it functions and may even expect the system to do the same. So it is not enough to be able to make inferences about people's needs and desires, we also need to anticipate the inferences that people will be making in turn.

Developing systems that can solve these problems is a daunting challenge, but fortunately, we can take inspiration from existing systems that must regularly interact with humans: other humans. And while nearly everyone has had to deal with other humans, cognitive science offers an extensive, systematic understanding of how we effectively solve the problem of understanding and interacting with others in our everyday lives.

In this section, we turn our focus towards what cognitive science has to say about people's inverse models of cognition and action. One of the most remarkable capacities that humans have is the ability to understand the hidden mental states that give rise to other people's observable behavior [53, 3]. This ability is often referred to as *theory of mind*, and cognitive scientists have studied it in adults [3], children [54], infants [55], and even other species [1]. Theory of mind has played a key role in our evolution as a social species capable of large-scale culture, coordination, and cooperation, and its development within the first year of life is a major milestone that enables us to comprehend and participate in the social world [2].

One exciting development in the recent study of theory of mind has been the application of ideas from economics, artificial intelligence, and control theory to characterizing mental state inference in computational terms. These approaches are reminiscent of methods familiar to engineers such as imitation learning, inverse reinforcement learning, and apprenticeship

learning [56, 57, 58]. However, in modeling the varieties of human social inference and interaction, they depart from and extend these ideas in numerous ways. This presents an exciting opportunity for collaboration between the cognitive sciences and engineering by providing new perspectives on inverse decision-making but in a shared conceptual and technical framework.

Here, we will focus on two lines of research on inverse models at the intersection of artificial intelligence and cognitive science. The first is work on cataloguing and systematizing the *conceptual primitives* involved in mental state inference—e.g., how people reason about mental entities like beliefs, desires, intentions, emotions, etc. The second is on inverse models in the context of teaching and communication, which are among the most basic types of social interactions that also expose the complexity of how humans use theory of mind productively. Along the way, we will discuss cases in which ideas from cognitive science have already been applied to the design of automated systems, limitations of existing approaches, and the possibilities for future research and applications.

3.1 Identifying the building blocks of theory of mind

Put in simple computational terms, theory of mind is an inference problem. That is, given limited observations of a process (e.g., a person's behavior), the task is to identify the hidden variables that produced those observations (e.g., the person's thoughts, desires, or feelings). Of course, this requires not only having concepts like *thoughts* and *desires* but an understanding of how these elements combine to produce behavior. This can be understood in rough analogy to another machine learning problem: parsing natural language. For example, inferring the *parse tree* of a particular sentence is jointly constrained by knowledge of *primitive types* of words (e.g., nouns, verbs, prepositions) and a *grammar* of how words tend to be combined. Recent models of theory of mind can be understood in terms of this linguistic metaphor: To explain how humans *parse* the behavior of other agents, we must understand the mental state primitives and mental state grammars that dictate how they are combined.

Incidentally, we have already covered one possible theory of how people parse behavior: expected utility theory [14, 15]. Taken as a generative model of people's intentional action, expected utility theory posits that others have beliefs about the state of the world (e.g., the belief that there is a burger joint down the street) and desires that certain states of the world are realized (e.g., the desire to eat a burger for lunch) and that people act rationally to realize their desires given their beliefs (e.g., the act of walking down the street to the burger joint). As an account of theory of mind, *inverse expected utility theory* makes several generalizable predictions that have been confirmed with human experiments. For example, adults, children, and infants can reason about how others integrate information about goals and action costs [4, 59, 60], features of different choices [61], the statistics of the environment [5], and limited perception of the environment [62]. Findings such as these have led to the proposal that human common sense psychology consists of a *naïve utility calculus* where we abstractly reason about other decision-makers as utility-maximizing agents [6].

At a broad level, the formal tools used characterize inverse utility theory will be familiar to those from a control background as they build on standard formalisms like MDPs, POMDPs, and inverse reinforcement learning [63]. However, there are interesting differences between how they are applied as models of human inference versus their typical engineering applications. For example, cognitive models tend to assume that people have highly structured representations of others' mental states (e.g., discrete objects), whereas in engineering applications the state representations are less structured (e.g., weights on a vector of continuous features [58]). This reflects the fact that cognitive scientists aim to explain how people can rapidly draw inferences based on only a few observations and a large base of background knowledge, while engineers are often trying to analyze large data sets of expert trajectories with minimal fine-tuning. As a result, cognitive scientists tend to use Bayesian methods while engineers are likely to be more familiar with methods designed to scale to large data sets. An important direction for future work is developing methods that cut across these different research agendas and can replicate the sophistication of human theory of mind with

tractable implementations.

Expected utility theory captures an important dimension of how humans parse others behavior in terms of beliefs, desires, and intentions. But, psychologists have also long studied other types of mental states that people reason about, such as emotions, habits, norms, rules, values, and social affinities, to name only a few [3]. While the traditional frameworks of expected utility and inverse reinforcement learning have not typically focused on these kinds of mental states, cognitive scientists have made great strides in extending the formalism to study these types of representations. For example, inverse planning models have been combined with models of habits [64], emotion and appraisal [65, 66, 67], responsibility judgments [68], values and norms in moral dilemmas [69], and social groups [70, 71]. Although do we do not typically think of robots as having these types of internal states, systems that are expected to interact with humans invariably need a basic understanding of the complete repertoire of psychological states that affect our individual and collective behavior.

The models described so far generally rely on assuming the standard formulation of rational action as optimizing a utility function defined over states of an MDP. For instance, they can express the idea of reaching a goal state while attempting to minimize costs along the way, but they cannot generally express history dependent or temporally specified constraints such as only going to a goal state only after accomplishing a subgoal. Recent work in both cognitive science and computer science have attempted to remedy this by introducing more flexible "logics" to express an agent's utilities, including linear temporal logic [72, 73, 74], finite state machines [75], and simple programs composed of sub-processes [76]. These methods are powerful because they can express rational cognition and action in a rich, compositional manner that may reflect human intuitions about agency. At the same time, this expressiveness comes at the cost of more complex and costly inference, and identifying appropriate constraints and settings in which tractable inference methods can be applied is an active area of research.

A complementary method for sidestepping strong assumptions about the structure of

rational action is to try and learn a theory of mind directly from data without recourse to pre-defined structural priors (e.g., using a neural network). Rabinowitz et al. [77] take this approach by generating a large data set of behaviors from synthetic agents with different goals, utilities, and perceptual abilities, and then using meta-learning to train networks with no prior conception of theory of mind or rationality to predict features of future behavior. Their networks were able to acquire a low dimensional embedding of behaviors and agent types capable of recreating several qualitative findings associated with theory of mind, such as inference about goals, costs, perceptions, and, to a certain extent, false beliefs. This work is an important demonstration of how relatively standard machine-learning methods can learn theory of mind-like representations given enough data and computation. However, studies by Nematzadeh et al. [78] have indicated that standard neural networks are limited in their capacity to explicitly represent false beliefs as they do not distinguish between appearance and reality, which is considered by psychologists to be a defining feature of theory of mind [54]. This suggests that some kind of structural priors about the nature of rational action are likely needed to capture the conceptual primitives and grammar comprising human theory of mind.

3.2 Identifying generalizable principles of communication and teaching

The previous section focused on theory of mind as a pure inference problem, but theory of mind also influences how people act and interact. Building on ideas originally explored by philosophers [79, 80, 81], developmental psychologists and linguists have extensively studied the principles underlying how people use theory of mind to learn from others and communicate [82, 2, 83]. A key idea is that when people communicate (e.g., by saying words, making gestures, providing examples, etc.), they do so with *communicative goals* like modifying the receiver's mental state or future actions. The person receiving these communicative signals can then reason about these goals to flexibly and efficiently draw inferences about what the

sender meant to convey. Note that this process requires both the receiver and sender to have a capacity for theory of mind and, more specifically, a capacity for *recursive theory of mind*, in which one agent reasons about another agent reasoning about the original agent (and potentially up to higher levels of recursion). Computational research over the past few years has formalized and extended many of these findings within the framework of probabilistic inference and decision-making [84, 85]. Here, we provide a broad overview of these developments.

One approach to studying communication in computational terms is the Bayesian pedagogy and cooperative communication framework [86, 84, 87], which characterizes how a teacher who presents data and a learner who interprets data should coordinate to efficiently and successfully communicate. To illustrate this idea, suppose you wanted to teach someone the concept of the even numbers by giving them a series of examples. Some examples will be better than others, for instance, {2, 2, 2} is technically a sequence of even numbers, but is not very informative, whereas {2, 4, 6} is clearly more helpful. Additionally, if the person receiving these examples knows you are being helpful, they can draw even stronger inferences based on what they are shown. Bayesian pedagogy models formalize this intuition about helpful, *informative* examples in terms of recursively defined teacher-learner equations, whose fixed points are optimal teacher-learner communication protocols. This approach has been successful in characterizing how both adults and children teach and learn concepts during pedagogical interactions [88, 89]. Additionally, recent theoretical work has established a direct correspondence between optimal transport problems and the equations that characterize teacher-learner fixed points [90, 91]. This opens the door for algorithmic insights to be shared between engineering disciplines such as operations research and the study of optimal cooperative communication in humans.

Cooperative communication models capture settings in which a teacher has a sole, explicit goal of being helpful and informative. However, communication is not always so clear-cut, and cognitive scientists are modeling more complex communicative situations within the

Rational Speech Act (RSA) framework [85]. For example, people may have goals other than being informative, like being succinct or inoffensive, and they may expect others to perform joint inference over these *non-communicative* goals. This process can give rise to linguistic phenomena as varied as hyperbole ("This kettle cost $1,000!" [92]) and politeness ("Your poem wasn't bad!" [93]). And while it may be extreme to expect automated systems to understand ironic humor, they will likely need to recognize more mundane forms of indirect speech.

More broadly, RSA can characterize how the context surrounding communicative acts shapes their meaning. For example, the statement "It's warm today" has different consequences if said during the spring versus the winter (e.g., you would likely wear shorts in the first but not the second case) [94]. Computational cognitive models of how humans flexibly reason about the shared context has been shown to be a key part of understanding general statements about the world [95]. Similarly, the ability to establish contexts as communicative (e.g., intentionally getting someone's attention in order to convey some information) has been shown to be an essential precursor to the types of cooperative communication interactions discussed above [96, 97]. Formal models that combine RSA with sequential decision-making and inference about whether partners have informative goals can provide a starting point for implementing such flexible reasoning about communicative context in autonomous systems [98, 99].

3.3 Applying human inverse models to the design of autonomous systems

Ideas from the computational cognitive science of mental state inference and communication have already begun to inform research in human-robot interaction and reinforcement learning. This showcases the potential for scaling up probabilistic models of cognition to complex, real-world domains, while also revealing novel insights about communication and teaching. For example, work in motion planning has led to robots capable of legible mo-

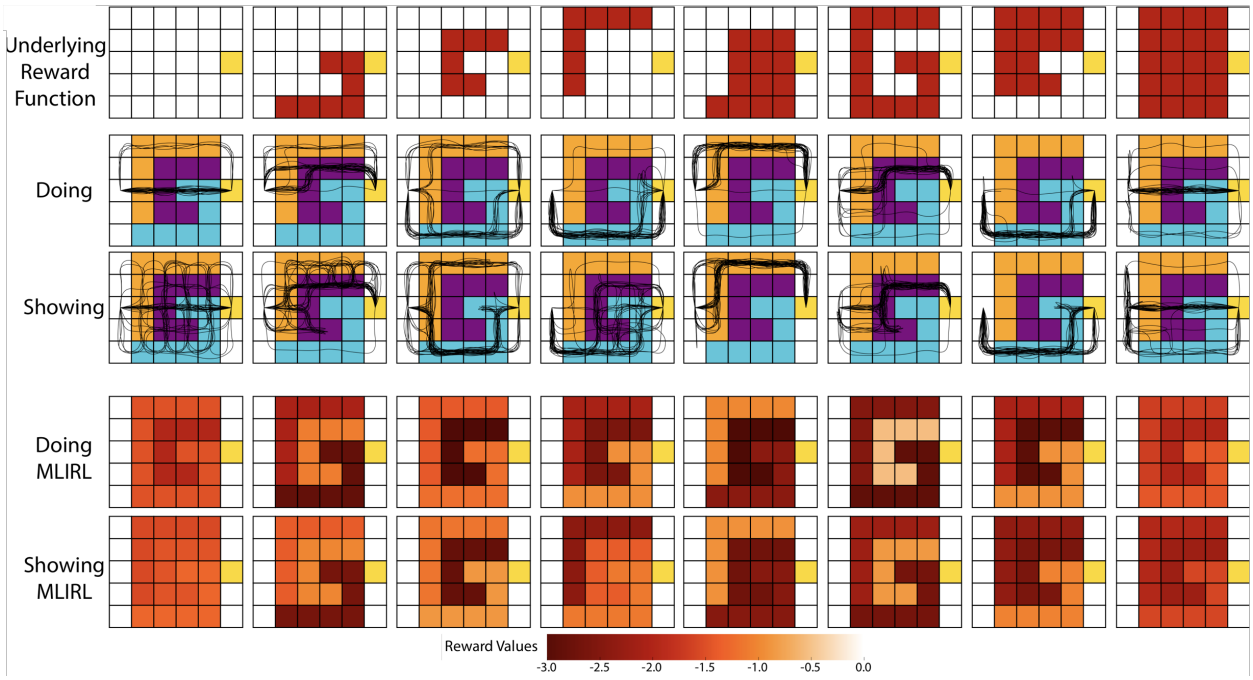

Figure 4: Learning from humans with communicative intent (data and figure from [10]). Ho et al. recruited human participants to perform grid navigation tasks that required reaching a goal state while not losing points. Each column represents one trial. Row 1 represents the true underlying reward function where white tiles are 0 points, red tiles are -2 points and the yellow goal tile is 10 points. Participants could not directly view the reward values, but were shown colors on the grid (orange, purple, blue) and told the value of each color (e.g., "orange and purple are safe"). Row 2 shows the visible layout of each grid and each black line represents one participant's trajectory on the task when they were only told to do the task. Row 3 shows the same grid and trajectories for participants told to do the task as well as *show* the reward function to an anonymous observer. Rows 3 and 4 show the reward weights estimated by Maximum-Likelihood Inverse Reinforcement Learning (MLIRL) [100] when given the *Doing* versus *Showing* demonstrations. Agents trained by Showing demonstrations obtain better estimates of the underlying reward function than those trained by Doing demonstrations.

tion [9, 11], in which action sequences are modified to allow humans to more quickly and successfully understand a robot's goals. Related work by Ho et al. [10] showed that when learning from human demonstrations, inverse reinforcement learning algorithms can benefit from being shown intentionally communicative expert behaviors (Figure 4). These modifications by both robots and humans are directly analogous to the approach taken in the Bayesian pedagogy framework, and demonstrate how consideration of communicative goals can facilitate human-machine teaching, learning, and cooperation.

Insights from cognitive science can also inform the design of learning algorithms themselves. For example, humans readily use rewards and punishments to modify the behavior of other animals and even other humans, and there is good reason to believe that similar principles of mental state inference and communication apply to these interactions [101]. In a series of experiments with humans interacting with different learning algorithms, Ho et al. [102, 103] demonstrated precisely this. Specifically, they found that people do not use rewards in a manner consistent with the standard interpretation as a quantity to directly maximize, as is typically done in reinforcement learning. Rather, they expect learners to reason about a teacher's pedagogical goals and interpret rewards as signaling information about whether an agent is "headed in the right direction" during the learning process. Such findings help motivate the development of learning algorithms that interpret reward in more sophisticated ways and attempt to infer people's teaching strategies and goals [104, 105].

Along similar lines, MacGlashan et al. [106] found that the structure of a human teacher's feedback depends on an agent's current stage of learning–a simple form of context. In a behavioral study, the authors had human participants interact with either a completely naïve agent or an expert agent that knew the optimal path to a goal. When the naive agent took a sub-optimal, but moderately good action, participants provided a high reward, while they gave the expert agent who should have known better a low or even negative reward for taking the same action. The logic of this strategy closely resembles that of the *advantage function*, which reflects the relative value of each action in a state under the agent's current

policy [107]. This insight motivated the design of the Convergent Actor-Critic by Humans (COACH) algorithm, which treats human feedback as an advantage signal. Subsequent work has also successfully applied COACH to training deep learning agents [108].

Additionally, researchers in control and robotics have developed unifying frameworks for modeling cooperative interactions likely to be faced in human-robot applications. For instance, cooperative inverse reinforcement learning [109] defines a general class of games from which specific cooperative strategies can be derived (e.g., legible motion, requests for information, etc.) based on a particular set up. Similarly, reward-rational learning [110] has been proposed as a framework for characterizing different types of human-robot interaction problems (e.g., learning from demonstrations, feedback, or examples) in terms of inference about human preferences that may be implicit. Combining these general computational frameworks with insights from the cognitive science of human decision-making will be an important direction for future research.

4 Opportunities for further research

The research we have reviewed is only a starting point for exploring potential applications of cognitive science to engineering, robotics, and control. There are a number of exciting directions based on these ideas we have covered. Here, we focus on three broad future directions.

Inverse resource rationality As we have noted, cognitive scientists have proposed that people reason about others as expected utility maximizers. Additionally, we discussed how expected utility theory is inaccurate and how resource rationality is a promising alternative framework. This raises the obvious possibility that people understand that others have limited cognitive resources and therefore reason about them not as pure utility maximizers, but as *resource-rational utility maximizers*. Importantly, progress here will depend on the continued development of plausible forward resource-rational models, especially models

of planning (e.g., [111, 112]). Nonetheless, several lines of research have already begun to explore inverse resource rationality. For example, research on perspective-taking and communication has shown that people can flexibly reason about the division of cognitive labor required for efficient communication [113]. Additionally, recent work has demonstrated how humans can infer preferences by jointly reasoning about the time it takes to make a decision and the decision itself [114], while other studies have shown how people reason about sub-optimal or inconsistent planning [115, 116, 117]. These ideas have begun to be incorporated into inverse reinforcement learning settings [118]. As with models based on inverse expected utility theory, understanding people's inverse models of resource-rational processes will be essential for how they interpret how machines make decisions and can also serve as inspiration for how machines interpret and interact with humans.

Blackbox vs. Theory-driven models for inferring intent In our discussion of forward and inverse models, we encountered three different examples of how machine learning tools can complement traditional psychological theory building. This included work on how humans make risky choices, work on how humans make moral decisions, and work on learning theory of mind concepts without hard-coding conceptual primitives or a principle of rationality. These approaches have illustrated how large data sets can be combined with machine learning tools to search the vast space of cognitive models in an efficient manner. However, they also illustrated some of the limitations of a purely blackbox approach and the benefits of also incorporating explicit, interpretable theories and structured prior knowledge. Thus, an important direction for future work will be developing methods that seamlessly integrate the benefits of each approach—on the one hand the scalability of blackbox methods, and on the other hand, the efficiency and interpretability of explicit psychological theories.

Schemas and mechanisms for human-robot interaction Unlike settings involving a single agent, interactions between two or more agents are challenging to design because they are fundamentally ill defined: Each agent might have their own goals and beliefs, which means there may not exist a single "yardstick" by which to measure whether the engineering

problem has been solved. This has occasionally prompted researchers to ask themselves "If multi-agent learning is the answer, what is the question?" [119].

We propose that cognitive science is uniquely positioned to provide guidance to question that multi-agent learning answers in the form of schemas and mechanisms grounded in the types of interactions that humans are adapted for. We have already encountered one example of this: The interaction of a teacher and learner engaged in cooperative communication can serve as a template for developing robots capable of legible action and value alignment [9, 109]. Beyond this, there are many other types of human interactions and socio-cognitive mechanisms that we have not discussed that could inspire future research. For example, humans form joint intentions to achieve shared goals, and this underlies our ability to cooperatively solve novel problems [120]. At a broader scale, human interactions are often shaped by norms, which can be understood as shared behavioral tendencies that generalize across interactions with different agents in a population. Norms and normative cognition have been extensively studied in cognitive science and psychology, and researchers have begun to explore these processes computationally [121]. Finally, an additional benefit of designing autonomous systems around how humans actually interact is the potential for new insights into those very interactions, leading to further collaboration between the cognitive sciences and engineering disciplines.

4.1 Conclusion

At the moment, there is no unified model of human cognition and decision-making that engineers can draw on when designing their systems. Nonetheless, cognitive science has much to offer those designing autonomous systems that interact with humans. In particular, cognitive science has a rich trove of theories and methods for systematically studying how humans think, decide, and interact with one another, and these discoveries are increasingly being couched in formal terms that are familiar to researchers in robotics and control. Here, we have discussed several recent frameworks and methodologies, such as the synthesis of

blackbox and theory-driven methods, resource-rational decision making, cooperative cognition, and rational speech act theory. We then surveyed how they have been applied to derive insights into the mechanisms and principles underlying people's forward and inverse models of decision making. As autonomous systems continue to become more commonplace in people's everyday lives, we expect engineers will also need to think systematically about how the humans their systems encounter will make decisions. We hope this review can provide clarity into the types of actionable insights and open questions that sit at the intersection of cognitive science and control research.

Acknowledgements: This work was funded by NSF grant #1545126, John Templeton Foundation grant #61454, and AFOSR grant # FA 9550-18-1-0077. Emojis in figures designed by OpenMoji, the open-source emoji and icon project. License: CC BY-SA 4.0.

References

[1] Premack D, Woodruff G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1(4):515–526

[2] Tomasello M, Carpenter M, Call J, Behne T, Moll H. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences* 28(5):675–690

[3] Malle BF. 2008. The fundamental tools, and possibly universals, of human social cognition. *Handbook of motivation and cognition across cultures* :267–296

[4] Baker CL, Saxe R, Tenenbaum JB. 2009. Action understanding as inverse planning. *Cognition* 113(3):329–349

[5] Lucas CG, Griffiths TL, Xu F, Fawcett C, Gopnik A, et al. 2014. The child as

econometrician: A rational model of preference understanding in children. *PLOS One* 9(3):e92160

[6] Jara-Ettinger J, Gweon H, Schulz LE, Tenenbaum JB. 2016. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences* 20(8):589–604

[7] Scassellati B. 2002. Theory of mind for a humanoid robot. *Autonomous Robots* 12(1):13–24

[8] Breazeal C. 2003. Toward sociable robots. *Robotics and autonomous systems* 42(3-4):167–175

[9] Dragan AD, Lee KC, Srinivasa SS. 2013. *Legibility and predictability of robot motion.* In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. IEEE

[10] Ho MK, Littman M, MacGlashan J, Cushman F, Austerweil JL. 2016. *Showing versus doing: Teaching by demonstration.* In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 3027–3035, pp. 3027–3035. Curran Associates, Inc.

[11] Fisac JF, Gates MA, Hamrick JB, Liu C, Hadfield-Menell D, et al. 2020. Pragmatic-pedagogic value alignment. In *Robotics Research*. Springer

[12] Sun R. 2008. *The Cambridge handbook of computational psychology.* Cambridge University Press

[13] Bernoulli D. 1738. Exposition of a new theory on the measurement of risk. *Econometrica* 22(1):22–36

[14] Von Neumann J, Morgenstern O. 1944. *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press

[15] Savage LJ. 1972. *The Foundations of Statistics*. Courier Corporation

[16] Ziebart BD, Maas A, Bagnell JA, Dey AK. 2008. *Maximum entropy inverse reinforcement learning*. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 3*, pp. 1433–1438

[17] McFadden D. 1973. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics* :105–135

[18] Tversky A, Kahneman D. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185(4157):1124–1131

[19] Peterson JC, Bourgin D, Agrawal M, Reichman D, Griffiths TL. 2021. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*

[20] Agrawal M, Peterson JC, Griffiths TL. 2020. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences* 117(16):8825–8835

[21] Griffiths TL, Lieder F, Goodman ND. 2015. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science* 7(2):217–229

[22] Lieder F, Griffiths TL. 2020. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* 43

[23] Kahneman D. 1979. Prospect theory: An analysis of decisions under risk. *Econometrica* 47:278

[24] Tversky A, Kahneman D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5(4):297–323

[25] Gigerenzer G, Todd PM. 1999. *Simple heuristics that make us smart*. Oxford University Press, USA

[26] Bordalo P, Gennaioli N, Shleifer A. 2012. Salience theory of choice under risk. *The Quarterly journal of economics* 127(3):1243–1285

[27] Ratcliff R, Smith PL, Brown SD, McKoon G. 2016. Diffusion decision model: Current issues and history. *Trends in cognitive sciences* 20(4):260–281

[28] Tsetsos K, Gao J, McClelland JL, Usher M. 2012. Using time-varying evidence to test models of decision dynamics: bounded diffusion vs. the leaky competing accumulator model. *Frontiers in neuroscience* 6:79

[29] Kiani R, Shadlen MN. 2009. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324(5928):759–764

[30] Latimer KW, Yates JL, Meister ML, Huk AC, Pillow JW. 2015. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* 349(6244):184–187

[31] Erev I, Ert E, Plonsky O, Cohen D, Cohen O. 2017. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review* 124(4):369–409

[32] Noti G, Levi E, Kolumbus Y, Daniely A. 2016. Behavior-based machine-learning: A hybrid approach for predicting human decision making. *arXiv preprint arXiv:1611.10228*

[33] Plonsky O, Erev I, Hazan T, Tennenholtz M. 2017. *Psychological forest: Predicting human behavior*. In *Thirty-First AAAI Conference on Artificial Intelligence*

[34] Geman S, Bienenstock E, Doursat R. 1992. Neural networks and the bias/variance dilemma. *Neural computation* 4(1):1–58

[35] Fudenberg D, Kleinberg J, Liang A, Mullainathan S. 2019. Measuring the completeness of theories. *arXiv preprint arXiv:1910.07022*

[36] Peysakhovich A, Naecker J. 2017. Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization* 133:373–384

[37] Plonsky O, Apel R, Ert E, Tennenholtz M, Bourgin D, et al. 2019. Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*

[38] Bourgin DD, Peterson JC, Reichman D, Russell SJ, Griffiths TL. 2019. *Cognitive model priors for predicting human decisions*. In *International Conference on Machine Learning*, pp. 5133–5141

[39] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, et al. 2018. The moral machine experiment. *Nature* 563(7729):59–64

[40] Thomson JJ. 1976. Killing, letting die, and the trolley problem. *The Monist* 59(2):204–217

[41] Simon HA. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1):99–118

[42] Horvitz EJ. 1987. *Reasoning about beliefs and actions under computational resource constraints*. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pp. 429–447

[43] Russell S, Wefald E. 1991. Principles of metareasoning. *Artificial Intelligence* 49(1-3):361–395

[44] Gershman SJ, Horvitz EJ, Tenenbaum JB. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349(6245)

[45] Lewis RL, Howes A, Singh S. 2014. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science* 6(2):279–311

[46] Lieder F, Griffiths TL, Hsu M. 2018. Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review* 125(1):1

[47] Hay N, Russell S, Tolpin D, Shimony S. 2012. Selecting computations: Theory and applications. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, ed. N de Freitas, K Murphy. Corvallis, OR: AUAI Press

[48] Callaway F, Rangel A, Griffiths TL. 2021. Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology* 17(3):e1008863

[49] Ortega DA, Braun PA. 2011. *Information, utility and bounded rationality.* In *International Conference on Artificial General Intelligence*, pp. 269–274. Springer

[50] Bhui R, Gershman SJ. 2018. Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review* 125(6):985

[51] Sims CA. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50(3):665–690

[52] Gershman SJ, Bhui R. 2020. Rationally inattentive intertemporal choice. *Nature communications* 11(1):1–8

[53] Gergely G, Csibra G. 2003. Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences* 7(7):287–292

[54] Flavell JH. 2004. Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly* 50(3):274–290

[55] Gergely G, Nádasdy Z, Csibra G, Bíró S. 1995. Taking the intentional stance at 12 months of age. *Cognition* 56(2):165–193

[56] Abbeel P, Ng AY. 2004. *Apprenticeship Learning via Inverse Reinforcement Learning*. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pp. 1–. New York, NY, USA: ACM

[57] Argall BD, Chernova S, Veloso M, Browning B. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57(5):469–483

[58] Arora S, Doshi P. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* :103500

[59] Jara-Ettinger J, Gweon H, Tenenbaum JB, Schulz LE. 2015. Children's understanding of the costs and rewards underlying rational action. *Cognition* 140:14–23

[60] Liu S, Ullman TD, Tenenbaum JB, Spelke ES. 2017. Ten-month-old infants infer the value of goals from the costs of actions. *Science* 358(6366):1038–1041

[61] Jern A, Lucas CG, Kemp C. 2017. People learn other people's preferences through inverse decision-making. *Cognition* 168:46–64

[62] Baker CL, Jara-Ettinger J, Saxe R, Tenenbaum JB. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1(4):1–10

[63] Jara-Ettinger J. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences* 29:105–110

[64] Gershman SJ, Gerstenberg T, Baker CL, Cushman FA. 2016. Plans, habits, and theory of mind. *PLOS One* 11(9):e0162246

[65] Ong DC, Zaki J, Goodman ND. 2015. Affective cognition: Exploring lay theories of emotion. *Cognition* 143:141–162

[66] Saxe R, Houlihan SD. 2017. Formalizing emotion concepts within a Bayesian model of theory of mind. *Current opinion in Psychology* 17:15–21

[67] Ong DC, Zaki J, Goodman ND. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science* 11(2):338–357

[68] Gerstenberg T, Ullman TD, Nagel J, Kleiman-Weiner M, Lagnado DA, Tenenbaum JB. 2018. Lucky or clever? from expectations to responsibility judgments. *Cognition* 177:122–141

[69] Kleiman-Weiner M, Gerstenberg T, Levine S, Tenenbaum JB. 2015. *Inference of Intention and Permissibility in Moral Decision Making.* In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, ed. D Noelle, R Dale, AS Warlaumont, J Yoshimi, T Matlock, CD Jennings, PP Maglio, pp. 920–925, pp. 920–925. Austin, TX: Cognitive Science Society

[70] Lau T, Pouncy HT, Gershman SJ, Cikara M. 2018. Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General* 147(12):1881

[71] Shum M, Kleiman-Weiner M, Littman ML, Tenenbaum JB. 2019. *Theory of minds: Understanding behavior in groups through inverse planning.* In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6163–6170

[72] Littman ML, Topcu U, Fu J, Isbell C, Wen M, MacGlashan J. 2017. Environment-independent task specifications via gltl. *arXiv preprint arXiv:1704.04341*

[73] Velez-Ginorio J, Siegel MH, Tenenbaum JB, Jara-Ettinger J. 2017. *Interpreting actions by attributing compositional desires.* In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, ed. G Gunzelmann, A Howes, T Tenbrink, J Davelaar. Cognitive Science Society

[74] Vazquez-Chanlatte M, Jha S, Tiwari A, Ho MK, Seshia S. 2018. Learning task specifications from demonstrations. *Advances in Neural Information Processing Systems* 31:5367–5377

[75] Icarte RT, Klassen T, Valenzano R, McIlraith S. 2018. *Using reward machines for high-level task specification and decomposition in reinforcement learning*. In *International Conference on Machine Learning*, pp. 2107–2116. PMLR

[76] Ho MK, Sanborn S, Callaway F, Bourgin D, Griffiths T. 2018. Human priors in hierarchical program induction. *Computational Cognitive Neuroscience (CCN)* 1

[77] Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SA, Botvinick M. 2018. *Machine theory of mind*. In *International conference on machine learning*, pp. 4218–4227. PMLR

[78] Nematzadeh A, Burns K, Grant E, Gopnik A, Griffiths T. 2018. *Evaluating Theory of Mind in Question Answering*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2392–2400. Brussels, Belgium: Association for Computational Linguistics

[79] Wittgenstein L. 1953. *Philosophical Investigations*. New York: MacMillan

[80] Grice HP. 1957. Meaning. *The Philosophical Review* 66(3):377–388

[81] Sperber D, Wilson D. 1986. *Relevance: Communication and Cognition*. Cambridge, MA, USA: Harvard University Press

[82] Clark HH. 1996. *Using language*. Cambridge: Cambridge University Press

[83] Csibra G, Gergely G. 2009. Natural pedagogy. *Trends in cognitive sciences* 13(4):148–153

[84] Shafto P, Goodman ND, Griffiths TL. 2014. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology* 71:55–89

[85] Goodman ND, Frank MC. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences* 20(11):818–829

[86] Shafto P, Goodman ND. 2008. *Teaching games: Statistical sampling assumptions for learning in pedagogical situations*. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society

[87] Landrum AR, Eaves Jr BS, Shafto P. 2015. Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences* 19(3):109–111

[88] Bonawitz E, Shafto P, Gweon H, Goodman ND, Spelke E, Schulz L. 2011. The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition* 120(3):322–330

[89] Bridgers S, Jara-Ettinger J, Gweon H. 2020. Young children consider the expected utility of others' learning to decide what to teach. *Nature human behaviour* 4(2):144–152

[90] Wang P, Wang J, Paranamana P, Shafto P. 2020. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems* 33

[91] Shafto P, Wang J, Wang P. 2021. Cooperative communication as belief transport. *Trends in Cognitive Sciences*

[92] Kao JT, Wu JY, Bergen L, Goodman ND. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33):12002–12007

[93] Yoon EJ, Tessler MH, Goodman ND, Frank MC. 2020. Polite speech emerges from competing social goals. *Open Mind* 4:71–87

[94] Tessler MH, Lopez-Brau M, Goodman ND. 2017. *Warm (for winter): Comparison class understanding in vague language*. In *15th International Conference on Cognitive Modeling*, pp. 193

[95] Tessler MH, Goodman ND. 2019. The language of generalization. *Psychological review* 126(3):395

[96] Csibra G. 2010. Recognizing communicative intentions in infancy. *Mind & Language* 25(2):141–168

[97] Scott-Phillips T. 2014. *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Macmillan International Higher Education

[98] Shafto P, Eaves B, Navarro DJ, Perfors A. 2012. Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental science* 15(3):436–447

[99] Ho MK, Cushman F, Littman ML, Austerweil JL. in press. Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*

[100] MacGlashan J, Littman ML. 2015. *Between imitation and intention learning*. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 3692–3698

[101] Ho MK, MacGlashan J, Littman ML, Cushman F. 2017. Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition* 167:91–106

[102] Ho MK, Littman ML, Cushman F, Austerweil JL. 2015. *Teaching with Rewards and Punishments: Reinforcement or Communication?* In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, ed. D Noelle, R Dale, AS Warlaumont, J Yoshimi, T Matlock, CD Jennings, PP Maglio, pp. 920–925, pp. 920–925. Austin, TX: Cognitive Science Society

[103] Ho MK, Cushman F, Littman ML, Austerweil JL. 2019. People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General* 148(3):520–549

[104] Loftin R, MacGlashan J, Peng B, Taylor M, Littman M, et al. 2014. *A strategy-aware*

technique for learning behaviors from discrete human feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28

[105] Hadfield-Menell D, Milli S, Abbeel P, Russell S, Dragan AD. 2017. *Inverse reward design*. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6768–6777

[106] MacGlashan J, Ho MK, Loftin R, Peng B, Wang G, et al. 2017. *Interactive learning from policy-dependent human feedback*. In *International Conference on Machine Learning*, pp. 2285–2294. PMLR.

[107] Baird L. 1995. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*. Elsevier

[108] Arumugam D, Lee JK, Saskin S, Littman ML. 2019. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*

[109] Hadfield-Menell D, Dragan A, Abbeel P, Russell S. 2016. *Cooperative inverse reinforcement learning*. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3916–3924

[110] Jeon HJ, Milli S, Dragan A. 2020. *Reward-rational (implicit) choice: A unifying formalism for reward learning*. In *Advances in Neural Information Processing Systems*, ed. H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin, pp. 4415–4426, vol. 33, pp. 4415–4426. Curran Associates, Inc.

[111] Correa CG, Ho MK, Callaway F, Griffiths TL. 2020. *Resource-rational Task Decomposition to Minimize Planning Costs*. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, ed. S Denison., M Mack, Y Xu, B Armstrong, pp. 2974–2980, pp. 2974–2980. Cognitive Science Society

[112] Ho MK, Abel D, Correa CG, Littman ML, Cohen JD, Griffiths TL. 2021. Control of mental representations in human planning. *arXiv preprint arXiv:2105.06948*

[113] Hawkins RD, Gweon H, Goodman ND. 2021. The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive science* 45(3):e12926

[114] Gates V, Callaway F, Ho MK, Griffiths TL. 2021. A rational model of people's inferences about others' preferences based on response times. *Cognition* 217:104885

[115] Evans O, Stuhlmüller A, Goodman N. 2016. *Learning the preferences of ignorant, inconsistent agents*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30

[116] Alanqary A, Lin GZ, Le J, Zhi-Xuan T, Mansinghka VK, Tenenbaum JB. 2021. Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*

[117] Berke M, Jara-Ettinger J. 2021. Thinking about thinking through inverse reasoning. *PsyArXiv preprint PsyArXiv:10.31234/osf.io/r25qn*

[118] Zhi-Xuan T, Mann J, Silver T, Tenenbaum J, Mansinghka V. 2020. Online bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems* 33

[119] Shoham Y, Powers R, Grenager T. 2007. If multi-agent learning is the answer, what is the question? *Artificial intelligence* 171(7):365–377

[120] Kleiman-Weiner M, Ho MK, Austerweil JL, Littman ML, Tenenbaum JB. 2016. *Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction*. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, ed.

A Papafragou, D Grodner, D Mirman, JC Trueswell, pp. 1679–1684, pp. 1679–1684. Austin, TX: Cognitive Science Society

[121] Hawkins RX, Goodman ND, Goldstone RL. 2019. The emergence of social norms and conventions. *Trends in cognitive sciences* 23(2):158–169