# Modeling, Replicating, and Predicting Human Behavior: A Survey

ANDREW FUCHS, Universitá di Pisa
ANDREA PASSARELLA and MARCO CONTI, National Research Council (CNR)

Given the popular presupposition of human reasoning as the standard for learning and decision making, there have been significant efforts and a growing trend in research to replicate these innate human abilities in artificial systems. As such, topics including Game Theory, Theory of Mind, and Machine Learning, among others, integrate concepts that are assumed components of human reasoning. These serve as techniques to replicate and understand the behaviors of humans. In addition, next-generation autonomous and adaptive systems will largely include AI agents and humans working together as teams. To make this possible, autonomous agents will require the ability to embed practical models of human behavior, allowing them not only to replicate human models as a technique to "learn" but also to understand the actions of users and anticipate their behavior, so as to truly operate in symbiosis with them. The main objective of this article is to provide a succinct yet systematic review of important approaches in two areas dealing with quantitative models of human behaviors. Specifically, we focus on (i) techniques that learn a model or policy of behavior through exploration and feedback, such as Reinforcement Learning, and (ii) directly model mechanisms of human reasoning, such as beliefs and bias, without necessarily learning via trial and error.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; *Theory of mind*; *Reasoning about belief and knowledge*; *Probabilistic reasoning*; *Temporal reasoning*; *Learning paradigms*; *Machine learning approaches*; *Knowledge representation and reasoning*; *Learning paradigms*; *Machine learning approaches*; • **Human-centered computing**;

Additional Key Words and Phrases: Artificial intelligence, Machine Learning, human behavior, cognition, bias, Human-AI Interaction, human-centric AI

**ACM Reference format:**
Andrew Fuchs, Andrea Passarella, and Marco Conti. 2023. Modeling, Replicating, and Predicting Human Behavior: A Survey. *ACM Trans. Autonom. Adapt. Syst.* 18, 2, Article 4 (May 2023), 47 pages.
https://doi.org/10.1145/3580492

This work was partly supported by the H2020 Humane-AI-Net project (grant No. 952026), CHIST-ERA (grant No. CHIST-ERA-19-XAI-010-MUR-484-22), and by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of partnership on "Artifical Intelligence: Foundational (PE0000013-program "FAIR").
Authors' addresses: A. Fuchs, Universitá di Pisa, Department of Computer Science and Institute for Informatics and Telematics (IIT), National Research Council (CNR), Via G. Moruzzi, 1, Pisa, Pisa, Italy, 56124; email: andrew.fuchs@phd.unipi.it; A. Passarella and M. Conti, Institute for Informatics and Telematics (IIT), National Research Council (CNR), Via G. Moruzzi, 1, Pisa, Pisa, Italy, 56124; emails: andrea.passarella@iit.cnr.it, marco.conti@iit.cnr.it.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2023 Copyright held by the owner/author(s).
1556-4665/2023/05-ART4 $15.00
https://doi.org/10.1145/3580492

## 1 INTRODUCTION

Future autonomous and adaptive systems are expected to further exploit the concept of cyber-physical convergence [25], and realize an environment where autonomous agents and humans work together as teams, understand each other, and anticipate each other's behavior and intentions. This is driven by the pervasive diffusion of connected devices in the physical environment, which are directly owned (e.g., personal devices) or in tight interaction (e.g., IoT devices) with the human user. The vast diffusion of AI can bring autonomy of agents to a new level, making their behavior much more refined and adaptive to the varying conditions of the environment and users. **Human-Centric AI (HCAI)** is expected to be a fundamental element in this vision. Quite interestingly, AI agents will need to interpret human behavior in context to better interact with users, understand their actions, predict their choices, and ultimately orchestrate between actions performed directly by humans and those delegated to AI agents in autonomy. To this end, it is essential to equip AI agents with practical models of human behavior.

Utilizing models of human behavior and decision making has spanned decades across numerous approaches and applications. Human behavior and reasoning enables complex behaviors and social structures. Consequently, these structures become multifaceted and grow significantly in complexity. Still, humans are generally quite successful at navigating this complex social structure. There are a multitude of attempts at explaining aspects of these capabilities, and this article discusses some of the more popular or persistent methods. The motivations behind research in the area of modeling human behavior and decisions are varied, so we limit the analysis to (a representative subset of) works providing *quantitative* models (e.g., math models or algorithms), as these are the approaches that allow one to "code" human behavior in autonomous systems. Some examples include better autonomous driving [39, 161], comprehension of mental states [17], and population-level modeling [66]. A primary distinction between the goals and approaches is often the fidelity of the replication and the expected deployment case. For instance, researchers may try to replicate the neurological pathways to mimic the neuro-physical process underpinning reasoning [11], or they may instead generate a computational model that is meant to mimic heuristically biased behavior [85]. In any case, a common aspect is the desire to use humans as the template for desirable patterns of reasoning.

Replicating or modeling human reasoning and abilities has motivated numerous research topics enabling autonomous and adaptive systems. These systems can be trained to work independently, in multi-agent systems, or in human-AI hybrid domains. In all of these cases, the resulting systems require the ability to both learn from the environment and adapt to observed changes. We focus primarily on topics relating to HCAI such as AI-assisted decisions or Hybrid Intelligence, which demonstrate cases where humans use or interact with AI systems. These interactions require differing assumptions regarding the dynamics between the human and the system, and how those dynamics impact the capabilities or characteristics of the systems. In a more direct case of **Human-AI Interaction (HAI)**, humans utilize the output of a system in their decision-making process. In this case, it is important the user clearly understands the output of an algorithm so they can effectively utilize the provided information. As an example, consider the use of AI or Machine Learning in medical diagnostics (i.e., HAI and AI-assisted decision making).

In what we consider one of the most relevant cases of HCAI for autonomous and adaptive systems (i.e., the case of Hybrid Intelligence), humans and AI systems form an interdependency. This can take multiple forms [53, 70] but leads to cases in which the human and the AI are expected to operate synergistically. Some examples include [33] Co-Evolution over Time, Human-in-the-Loop Learning, Interactive Learning or **Active Learning (AL)**, and Socio-Technological Ensemble. Given these paradigms, the human and AI can be paired in multiple configurations depending on technical, social, and related considerations (see the work of Wilkens et al. [160] for more details).

Although emerging, the literature on HCAI and related human behavioral models is quite vast already, and it is apparent that various types of HCAI exist. As an initial example, consider HCAI systems that focus primarily on explainability [37, 58, 60, 61, 164]. Approaches for explainability serve to make complex algorithms and methods explainable to humans. As an example, consider the preceding case of AI-assisted decisions in medical diagnostics. The system must be effective and demonstrate high accuracy, but must also prove worth using from a user's perspective. If not, the utility of the system may go unnoticed and unused. The use of explanation with respect to the decision or underlying algorithm improves user awareness as it aids their determination regarding the trustworthiness of the system and its output. A system demonstrating trustworthiness will of course have a higher likelihood of use. A further form of HCAI (typically referred to as Hybrid Intelligence) consists of AI agents and humans interacting directly, and impacting each other's operations [53, 70]. Examples of such cases include AI agents learning in the presence of human teachers [59, 87, 98, 100, 113, 118], or humans exploiting the outcome of AI agents to acquire better comprehension of a phenomenon [131]. Moreover, humans and AI agents may perform a common operation as a team [96, 120]. In this context, designing approaches to orchestrate delegation of tasks and decisions among the team's members is also fundamental [101, 116].

The literature we consider on modeling HCAI systems for achieving higher levels of autonomy and adaptation can be classified as shown in Figure 1. Specifically, we will discuss some popular topics and applications relating to HCAI, HAI, and Hybrid Intelligence. These topics represent methods that attempt a model mimicking or inspired by various biological/neurological, cognitive, and social levels of reasoning. Additionally, we discuss how these topics align with application areas of interest. These application areas cover a wide assortment of both scenario and level of autonomy expected—for instance, demographic preferences [66] or something as safety-critical as fully autonomous driving [39, 161]. The specific scenario can rely significantly on the level of autonomy expected and the level of risk or autonomous control humans are willing to accept. For these topics, we provide underlying principles and definitions, relevant examples of use cases and their approaches, and further examples of relevant survey/review papers and related resources. Finally, we provide additional details regarding some common application areas for these topics.

The rest of the article is structured as follows. In Section 2, we present common concepts and key application areas. In Section 3, we discuss learning methods that generate a model of behavior either by trial and error or by learning from observations of others. These techniques learn from feedback denoting desirable behavior and adapt their policy according to this feedback. Next, in Section 4, we discuss methods that attempt to model the mental states of others, utilize world models to simulate human knowledge, or discuss bias and fairness in representations and reasoning. In Section 5, we discuss methods that attempt to replicate the cognitive abilities and suboptimality found in humans. These allow for models that replicate biased or bounded rational behavior inspired by human cognition. A key aspect is modeling human cognitive resources and techniques adopted by humans to efficiently use them, possibly at the cost of obtaining imprecise understanding of the learned process or introducing mistakes. In Section 6, we focus on approaches that model uncertainty in the human reasoning process, including cases where choices do not appear to be the outcome of a rational process. Finally, Section 7 provides a brief overall comparison of the considered approaches.

We will not provide a comprehensive coverage of all aspects relating to the listed topics, but instead attempt to provide a useful assortment as a demonstration of topics of interest offering potential areas for further exploration. Each section describing a topic is organized according to a common structure. First, we point to specific surveys and related work dealing in greater detail with that topic. Then, we discuss the general principles. Next, we discuss one concrete example where those principles are made practical. Finally, we briefly mention additional examples where the same principles have been applied.


Fig. 1. Taxonomy of concepts.

## 2 HCAI ORTHOGONAL CONCEPTS AND SAMPLES OF APPLICATION AREAS

In this section, we discuss some common concepts related to HCAI that cut across the various modeling methods presented in detail in the rest of the article. Moreover, we will briefly discuss popular application areas demonstrating uses of techniques discussed. This list is not comprehensive but serves to demonstrate topics that are likely more familiar and of immediate interest. The approaches used demonstrate methods that serve to replicate, model, or learn from human behavior and capabilities.

### 2.1 Orthogonal HCAI Concepts

It is important to note that different issues and considerations arise from varying requirements—for instance, ensuring the AI system is not difficult to use or explain to users, difficult to manage or maintain, or perceived as creepy by the potential users [37, 164]. Additionally, the systems

will likely require awareness and capabilities supporting numerous types of intelligence (social, emotional, physical, etc.) to best understand and interact with humans [24] while operating autonomously. Further, the reliability, correctness, and resulting impact of the system can be viewed as an important factor [110].

A key aspect in human-centric paradigms such as HAI or Hybrid Intelligence is that the behavior of the human and the AI are assumed to impact each other [117]. In the case of Hybrid Intelligence, each side is providing a deeper aspect to the relationship. In general, this would require AI systems that can observe and understand humans to improve their behavior in this hybrid domain [71] via adaptability. As an example, in some cases of AL or **Reinforcement Learning (RL)**, a model is being learned with the help of, or in observation of, a human teacher [59, 87, 98, 100, 113, 118]. In such a case, we are expecting the AI system to continually refine both its model relative to the data samples while also improving its ability to know when to ask for help. Further, one could expect the human to use their understanding of the system implicitly or explicitly to improve the utility of the samples based on the observed progression of the system [131]. In some cases, the samples or information can come as human retellings of past experience [79]. This can also be reversed in the sense that the system can be designed to improve the types of responses to guide the human and improve the information received or queries of the user [152].

Systems can also be designed to learn to work in tandem with the human as a team. In such cases, it is often desired to have the system augment the abilities of the human or maintain autonomous control over an aspect of the task that is more challenging for the human (and is less critical with respect to the larger goal). For instance, AI systems can be trained to assist the human in a navigation task [120]. In such a case, there are aspects of the problem that are much easier for either the human or the AI, so sharing the responsibilities allows for improved performance over either working independently. For more direct assistance to the human, in the work of Morrison et al. [96] we see an AI system augmenting the senses of the user to assist visually impaired users in a social context. This demonstrates a system more closely integrated into the sensory systems of the human and is intended as an unobtrusive augmentation of their sensory capabilities. From another perspective, the goal could be a system that can operate as an independent agent in the environment [157]. In these cases, the agent is expected to respond to the observed behavior of the human to assist or avoid interfering as each work to achieve their tasks. In either case, it is important to consider how the two (or more) are expected to interact and respond to observed behavior [43, 168].

Another important aspect of Hybrid Intelligence is delegation (how and when to delegate) and which tasks can be performed without human intervention [101, 116]. There needs to be graceful handling in the event the human and AI system are both attempting to control the system. Additionally, there needs to be clear guidance regarding when either should be in control. For instance, it is important to understand the reliability of the system and where it is likely to encounter errors. Beyond simply impacting performance, system errors can impact the human's perception of the system and its overall utility. Consequently, this can impact the interactions between the system and the human user [14].

To facilitate understanding and learning for the AI systems, it is important to support methods for representing the observed behavior from the human in a manner that can be tractable and potentially simplify learning. As an example, Xie et al. [163] encode the observations in a latent space and then learn a model of behavior corresponding to the latent representation. Such an encoding, among other benefits, can allow the agent to abstract the observations to support connections between similar observations. Additionally, systems require the means to observe and adapt to humans [113, 130]. An important example is human-aware robot navigation [91, 93] and human-robot interaction scenarios [132]. In the navigation context, the motion, goals, and general

behavior are crucial for the robot to successfully navigate the environment. Similarly, models can be generated to estimate or predict the feedback expected from a human collaborator or teacher to boost training of the AI system [99, 100].

## 2.2 Application Areas

*2.2.1 Robotics.* Robotics can utilize human demonstrations to learn skills or behavioral policies. For instance, humans can provide demonstrations for a robot learner using a policy-gradient RL policy, which can then be improved through practice and exploration by the learner [5]. Similarly, humans demonstrate the ability to adapt skills to variations and new scenarios. This adaptability has been explored in RL agents in robotics to enable generalization from initial demonstrations or exploration [68, 118], which enables aspects of learning from demonstrations and an interactive learning process as seen in **Imitation Learning (IL)**, **Inverse Reinforcement Learning (IRL)**, and AL. Robot agents can also learn to anticipate the movement of humans or other artificial agents to compensate for their movements or rendezvous at a later position [91, 93, 156], and this concept can be extended to further topics in human-robot collaboration [132]. Additionally, robot vision can be designed to replicate models of human vision mimicking foveation [15], which allows agents to observe visual stimuli with similar attention and focus to stimuli. These topics demonstrate how robots can be designed to learn from humans and also learn to operate in an environment alongside humans. The approaches for these solutions span multiple disciplines, including ones described in this article (RL, IRL, etc.).

*2.2.2 Driver Prediction and Autonomous Driving.* There have been numerous examples of research performed to model and predict behavior in a driving scenario [77]. These topics include methods that model pedestrian behavior [23] or the behavior of other drivers [16, 22, 39]. This manner of modeling serves to train autonomous vehicles how to successfully drive while compensating for the behaviors of others through the use of models based on IRL IL, and related methods. This predictive power is essential so systems can anticipate and react to the non-uniform nature of human behavior. In addition to modeling the behavior of other drivers, systems have been investigated that model the vehicle control behavior of drivers [109]. This allows for comprehension and modeling of driver control movements when operating a vehicle. For example, there is a need to model and predict the outcome of control delegation between the autonomous system and the human driver. To do so, researchers have investigated cognitive models to predict the time to take over control given the type and difficulty of actions the human driver is performing when the control is switched [88, 128, 129]. This allows for a model that can simulate the cognitive and bio-mechanical responses when changing tasks for the human drivers.

*2.2.3 AI in Games and Teaching.* In the area of video and serious games, AI is considered with respect to multiple aspects. In one aspect, researchers and developers are investigating techniques to integrate AI into game development and game character behavior [18, 162, 166, 167, 169]. This allows for levels, players, and so forth to be generated or controlled by adaptive behavior models to create a broader scope of experiences. These examples demonstrate using approaches such as planners (see Section 4), RL, and those related to improve the performance, adaptability, or creation of games. Non-player character behavior can be supported by these techniques to generate policies based on historical gameplay data or learned models of play. This can be from learned models of behavior or based on past human player choices [31]. Non-player characters with these characteristics could prove better suited to respond to player choices and allow for more outcomes. Similarly, the generation of levels and scenarios can also be expanded by allowing for more dynamic combinations of resources by the system. In the case of serious games, systems can be implemented to train human users and learn models of their behavior [74]. This allows

for systems that can teach and allow for improvement with a more thorough understanding of the user's behavior. For instance, cognitive architectures (see Section 5.2) can be used to define intelligent tutoring systems [139].

*2.2.4 Agent-Based Modeling.* To achieve a model of human behavior and decision-making, numerous topics have been investigated. Some examples include **Agent-Based Modeling (ABM)** [35, 52, 66, 73, 136], which allows for models of groups of people or populations. In ABM, human behavior models are often defined and then studied in an environment over a simulated timeline. The agents follow the defined patterns of behavior and the resulting global patterns can be analyzed. For instance, population segregation based on demographic preferences regarding neighbors can be modeled by defining a diversity preference and modeling the movement of agents in an environment [66]. Populations of humans can also be modeled with other techniques designed to model the interdependence of the agents. The interactions can be modeled mathematically with humans represented as nodes in a network, particles in an environment, or more [36]. Additionally, the model of human behavior in ABM can also be supported by models of uncertainty, bounded rationality, or cognitive architectures, which we discuss in Section 5 and Section 6. As we will focus on more direct and individual-level models of behavior, we will not present this topic in greater detail in this article.

## 3 LEARNING HUMAN BEHAVIORS BY EXPERIENCE AND FEEDBACK

In the following sections, we discuss methods that learn patterns of behavior by exploration and observation of feedback. In Section 3.1, we discuss learning agents that perform RL by observing rewards denoting the desirability of actions in a particular state or context. Extending this concept of feedback-based learning, we discuss IRL and IL in Section 3.2. These learning agents develop a model of behavior attempting to replicate the observed behavior of others. In IRL, the agent generates an estimate of the feedback that generated the observed behavior and learns a corresponding model of behavior. A similar attempt to replicate observed behavior is seen in IL. In this case, the learner does not estimate the feedback signal but instead directly learns a model of behavior from the observations. The final section, Section 3.3, discusses AL—a paradigm in which the agent is able to query a teacher for feedback regarding a subset of input values. This allows learning based on an estimated confidence regarding inputs and to utilizing teacher expertise to update knowledge.

### 3.1 Reinforcement Learning

*3.1.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [83, 112, 130].

*3.1.2 Principles and Definitions.* RL is a method by which situations are mapped to actions to maximize a reward signal [142]. The maximization is performed by the learning agent through exploration of an environment and the possible actions. This exploration generates a feedback signal via rewards that the agent uses to learn the behaviors resulting in the most desirable feedback. The feedback received is provided immediately or can be a result of a sequence of actions. For example, an agent could receive a reward for each step they make in an environment or simply a single reward at the end of a training session that corresponds to the outcome. The different parameters of the problems lead to numerous techniques for learning optimal behavior policies.

*Markov Decision Process.* In RL, agents are attempting to solve a sequential decision process, which is represented by a **Markov Decision Process (MDP)**. This representation allows a scenario to be formally modeled with underlying assumptions, such as dependence on past states. An MDP can be represented by the tuple $\{S, A, T, R, p(s_0), \gamma\}$ [114]. $S$ refers to the states of the

environment that are traversed by executing actions from $A$. The transition between states follows the transition probabilities $T : S \times A \rightarrow p(S)$. As a means for feedback, the reward function $R : S \times A \rightarrow \mathbb{R}$ provides reward signals based on the selected action. Note that $R$ can also be defined with the inclusion of the resulting state $R : S \times A \times S \rightarrow \mathbb{R}$. Additionally, $p(s_0)$ defines the probabilities over initial states and $\gamma$ defines the discount parameter (defined in Section 3.1.3).

*3.1.3 Policies and Learning.* Given a scenario or MDP, an agent is trained to find a policy $\pi : S \rightarrow p(A)$ that defines a likelihood of actions given current state. An agent's policy is learned through trial and error by exploring the given MDP and observing the rewards $r \in R$. A policy is learned via an estimate of action utility based on past observations and estimates of trajectories. The agent can generate an estimate of discounted return based on a discounted sum of future rewards:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{1}$$

where $R_{t+k}$ denotes the reward observed $k$ timesteps after time $t$ and $\gamma$ is the discount parameter. This describes a measure of estimated return based on observations of trajectories. The value of $\gamma$ determines how quickly the scale of future rewards decays, which impacts how strongly those observations impact the estimates. With this method of estimating returns, agents can generate a model of likely utility for states or state-action pairs. This concept is used to define a value function $v(s)$ given state $s$ with the assumption that the agent starts in state $s$ and executes actions according to their policy $\pi$ in subsequent states. The distinction being that the estimate is based on behavior determined by a policy $\pi$. This means that the estimated value will consider the estimates of future values given the current state and expected trajectory. In the work of Sutton and Barto [142], $v(s)$ is defined as

$$v_\pi(s) = \mathbb{E}_\pi \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s \right], \forall s \in S. \tag{2}$$

Similarly, the action value function is used to estimate the value of executing an action $a$ in a given state $s$:

$$q_\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s, a_t = a \right], \forall s \in S. \tag{3}$$

These are fundamental equations with respect to behavior policy learning via RL. Various methods are used to learn a representation of value following Equation (3). This is done by utilizing the trial-and-error learning process of RL. Agents explore the environment and select actions (following a learning scheme—e.g., Epsilon Greedy) to observe the effect of actions. The effect is reflected in the rewards observed, which are used in the estimated state-action value function to incrementally improve the estimated utility of actions. This is done by refining the estimate through a modeling process that uses immediate rewards and the current estimate of value to refine the estimate. This cycle is used to learn a policy through the feedback loop created by taking an action, observing an outcome, and updating the policy of actions accordingly. As an example, a temporal difference method can be used in Q-Learning to learn a state-action value function:

$$Q_{t+1}(s, a) = (1 - \alpha) Q_t(s, a) + \alpha [r + \gamma \max_{a'} Q_t(s', a')], \tag{4}$$

where $\alpha$ is the learning rate used to discount the scale of the current estimate in the update to the estimated value. As can be seen, there is a recursive relationship between the current state's utility and the value of future states in the discounted $\gamma \max_{a'} Q_t(s', a')$ term. This enables the relationship between the value of the current state and the value of future states under the current policy.

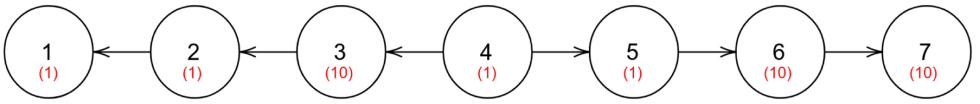


Fig. 2. Sample MDP showing issue with myopic view. State values are in parentheses below state labels.

In RL, the method for defining and finding optimal behavior depends on the nature of the elements of the observation available to the agent. In the most general case, such as Q-Learning demonstrated earlier, the agent observes the current state and selects an action according to $\pi$. Typically, this is done by identifying $\text{argmax}_a\, q_\pi(s, a)$. An important aspect of basing decisions on discounted future utilities is the possibility of MDPs that could result in suboptimal behavior given a myopic view. For instance, Figure 2 provides an example of how too greedy a nearsighted view of state values would lead an agent to bias its behavior to lead from state 4 to state 1 with a sum of rewards equal to 12 rather than the optimal choice, which is 4 to 7 with a sum of rewards equal to 21. Such a scenario can be quite common and is one of the many challenges for successful policy learning in RL.

RL contains numerous examples of constraints that differentiate families of problems. For instance, it is possible to have an environment that is only partially observable [134]. In this case, not all aspects of a state are observable to an agent, so there is uncertainty regarding the exact state the agent is occupying. This results in a need for the agent to estimate the current state and use that estimation in the updates to the value function. In another case, the actions may not lead to an outcome with 100% certainty. An example is a stochastic gridworld problem (e.g., OpenAI Gym Frozen Lake [20]). In this scenario, agents try to navigate a 2D world by moving up, down, left, or right. When an agent selects an action, there is a non-zero probability that the agent moves to an unintended state. For example, an attempted "up" action could in fact move the agent to the left. In this case, the learning method must support this uncertainty. It is worth noting that the previous examples represent scenarios supported by tabular methods. The use of RL has also been extended to continuous cases with the help of Deep Learning. This is referred to as Deep RL and is commonly used in the case of continuous and/or large domains such as robotic control or video games [10].

*3.1.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [41].

### 3.2 Inverse Reinforcement Learning and Imitation Learning

*3.2.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [9, 39, 65, 170].

*3.2.2 Principles and Definitions.*

*Inverse Reinforcement Learning.* IRL is a method by which an agent learns from examples of behavior without access to the underlying reward function motivating the behavior. The key distinction being that the agent is trying to replicate or approximate the reward function $R_E$ or policy $\pi_E$ that caused the exemplar behavior. This results in effectively needing to learn a reward while simultaneously attempting to learn optimal behavior policy under the current estimated reward function. As such, the agent is performing two interdependent tasks. Given a policy $\pi_E$ or a set of $N$ demonstrated trajectories

$$\mathcal{D} = \{\langle (s_0, a_0), (s_1, a_1), \ldots, (s_j, a_j)\rangle_{i=1}^{N} : s_j \in S; a_j \in A; i, j, N \in \mathbb{N}\},$$

the agent is tasked with learning a representation that could explain the observed behavior [9].

Generally speaking, there are numerous methods or approaches with respect to IRL. Therefore, we will be unable to address all the techniques in this section. We will instead provide some preliminary examples for intuition regarding common techniques and underlying principles. One method for the IRL task is that of apprenticeship learning [1]. In this case, there is an assumed vector of state-related features $\phi : S \rightarrow [0, 1]^k$ that support the reward $R^*(s) = w^* \cdot \phi(s)$, with weight vector $w^*$. The feature vectors $\phi$ refer to observational data corresponding to the states (e.g., a collision detected flag). Given the definition of reward, the value of a policy $\pi$ can be measured by

$$\mathbb{E}_{s_0 \sim D}[V^\pi(s_0)] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)|\pi\right] = w \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t)|\pi\right] \tag{5}$$

with the initial states being drawn $s_0 \sim D$ and with behavior following from the policy $\pi$. Then, the *feature expectation* can be defined as

$$\mu(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t)|\pi\right], \tag{6}$$

which is used to define a policy's value $\mathbb{E}_{s_0 \sim D}[V^\pi(s_0)] = w \cdot \mu(\pi)$. Given the estimation of feature expectation $\mu(\pi)$, the goal is to find a policy $\tilde{\pi}$ that best matches the observed demonstrations. To do so, this requires a comparison between $\tilde{\pi}$ and $\pi_E$. Since the policy $\pi_E$ is typically not provided, an estimate $\hat{\mu}_E$ based on demonstrations is needed. This can be accomplished by an empirical estimate:

$$\hat{\mu}_E = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{\infty} \gamma^t \phi\left(s_t^{(i)}\right), \tag{7}$$

for a given set of trajectories $\{s_0^{(i)}, s_1^{(i)}, \dots\}_{i=1}^m$.

Given this structure, algorithms can be defined for iteratively refining the weight vectors and resulting policies. This enables the two components of the IRL paradigm. First, the estimated weight vectors $w^{(i)}$ define an estimated reward function that can be refined by updating the weight vector to reduce the discrepancy between the two estimated feature expectations $\hat{\mu}_E$ and $\mu(\pi)$. Second, the estimated reward with the current weight $w^{(i)}$ enables learning a policy using RL that is optimal according to Equation (5). This is used in the first step to measure value discrepancy. The process provides a cyclical method of reward refinement and policy learning.

It is worth noting that needing to find a suitable reward function and a policy increases the complexity of the problem. The cycle of improving the reward requires the agent to retrain a behavior policy that reflects the new reward. However, finding an accurate representation of the reward would afford the agent a more general understanding of the behavior. Having access to a reward function allows an agent to understand desirable behavior at an abstract enough level to potentially transfer its understanding to a new environment. Of course, if the approximated reward is not accurate enough, one would expect this to result in issues. In any case, there is the potential for increased generality of the resulting agent. Finally, note that an exact replication of the underlying reward is not entirely necessary. In fact, an affine transformation of the true reward function would result in an equivalent policy [9].

*Imitation Learning.* IL is a process that attempts to learn a pattern of behavior under a given task [106] based on given expert demonstrations. At an abstract level, IL is closely related to IRL. In both cases, the goal is to utilize observations of behavior to train a behavioral model or policy. The key distinction comes in the structure of the learning process. In IRL, the learning process attempts to learn a suitable reward function that aligns with the demonstrated behavior. The learning process then uses the learned reward to generate a policy of behavior. For IL, the learning process is not

designed to generate a reward function that fits the demonstrations; instead, the process attempts to directly learn a model of behavior best fitting the demonstrations.

As a demonstration of the principles, IL can be formulated as follows. Given a trajectory $\tau = [\phi_0, \ldots, \phi_T]$ of features $\phi$, the learner is tasked with learning a policy reproducing the behavior. The vectors $\phi$ represent the features of the environment or system at each stage of the trajectory. The context, or state, $s$ representing the system state can be used in conjunction with an optional reward parameter $r$ as components of the underlying optimization problem.

With a set of demonstrations $\mathcal{D} = \{(\tau_i, s_i, r_i)\}_{i=1}^{N}$ of trajectories $\tau$, contexts $s$, and rewards $r$, (Note: the reward values $r_i$ might not be provided), the goal is to find a policy $\pi^*$. The policy should minimize the discrepancy between the distribution of features from the expert $q(\phi)$ and the distribution of features from the learner $p(\phi)$:

$$\pi^* = \operatorname{argmin} D(q(\phi), p(\phi)), \tag{8}$$

where $D$ is a discrepancy measure such as Kullback-Leibler. A key feature of Equation (8) is that it promotes the alignment of behavior through direct observation by penalizing distributions diverging from the demonstration. This goal of low discrepancy guides the policy search toward those policies that best fit the observed distribution over features. Overall, the method for generating policies will vary with each approach, but the underlying principle of direct replication is still a key component.

*3.2.3 Applications and Recent Results.* To demonstrate the preceding concepts, we present an IRL example [165] in which an agent learns a model to replicate human gaze patterns when searching for a visual target. The proposed approach, referred to as Dynamic-Contextual-Belief (DCB), uses imagery data with human-annotated gaze fixations. Further, the: approach uses simulated fovea to learn a plausible model of desirable objects for attention. This is used to model how a human's gaze shifts around an image while attempting to locate an object. The approach showed an improved ability to identify particular objects of interest over generating a saliency map to demonstrate attention.

DCB is composed of three components: fovea, contextual beliefs, and dynamics. The fovea serves to mimic human vision and provide a sub-region in high detail while blurring or masking the remaining region. The masking selects a sub-region of the image to represent in high resolution, whereas the remaining regions are represented using a blurred version of the image. This approximates the effect of the fovea on human vision and represents the fixation of the observer. The contextual beliefs represent a person's understanding of the contents of the scene, such as objects and background items of an image. Last, the dynamics component collects information regarding the focal fixations during search. The dynamics are represented as a transition between versions of the image, which are updated based on iterations of fixation. Each region that receives fixation is replaced by its high-resolution representation, resulting in a transition represented by

$$B_0 = L; B_{t+1} = M_t \odot H + (1 - M_t) \odot B_t, \tag{9}$$

where $B_t$ represents the belief state after $t$ fixations and $M_t$ is the mask generated by the $t^{th}$ fixation. $L$ and $H$ are belief maps that represent object and background locations for low-resolution and high-resolution images, respectively. Based on this definition, we can see that the representation of the image and the beliefs $B_t$ regarding contextual information and item locations are updated based on the iterative search conducted by moving the fixation around the image.

To represent the reward and learn a policy representing visual search behavior, Generative Adversarial Imitation Learning (GAIL) is utilized, which is a framework that utilizes the adversarial paradigm, with networks representing the discriminator $D : S \times A \rightarrow (0, 1)$ and generator $G$. These are used to train the system to generate data matching the patterns of the original sampled data.

The discriminator is tasked with learning to distinguish between real human data and those generated artificially by $G$. The generator $G$ is tasked with learning to generate data convincing enough to the discriminator $D$ to be labeled as real rather than artificially generated. This is accomplished by maximizing an objective function:

$$\mathcal{L}_D = \mathbb{E}_r[\log(D(S,a))] + \mathbb{E}_f[\log(1 - D(S,a))] - \gamma\mathbb{E}_r[||\nabla D(S,a)||^2], \tag{10}$$

where $D(S,a)$ is the output of $D$ given the state-action pair $(S,a)$, $\mathbb{E}_r$ refers to the expectation over real state-action pairs, and $\mathbb{E}_f$ refers to the expectation over fake search transition samples from $G$. The gradient term at the end of Equation (10) serves to improve the convergence rate. The definition of the reward is based on the output of the discriminator:

$$r(S,a) = \log(D(S,a)). \tag{11}$$

For the generator, the performance uses Equation (11) and is tasked with maximizing

$$\mathcal{L}_G = \mathbb{E}_f[\log(D(S,a))] = \mathbb{E}_f[r(S,a)], \tag{12}$$

which shows that the higher likelihood of being real the discriminator places on generated data, the higher the resulting reward will be. To find an RL policy for the generator, Proximal Policy Optimization is utilized with the following representation:

$$\mathcal{L}_\pi = \mathbb{E}_\pi[\log(\pi(a|S))A(S,a)] + H(\pi), \tag{13}$$

where the advantage function $A$ is estimated using Generalized Advantage Estimation (GAE). The advantage represents the gain observed by taking action $a$ versus the policy's default behavior. This definition of loss for the policy learning helps guide the learner toward actions that will result in higher advantage over other actions. The term $H$ is the max-entropy IRL term $H(\pi) = -\mathbb{E}_\pi[\log(\pi(a|S))]$, which helps improve convergence.

To test the approach, images from the COCO-Search18 dataset were included based on relevant selection criteria. The algorithm was tested against relevant baselines and against human performance. In the tests, results were used from 10 participants who viewed the 6,202 images in the dataset. For the participants, eye tracking was performed during the search task. The algorithms compared against were a random scanpath, a ConvNet detector, Fixation heuristics, a Behavior Cloning CNN, and a Behavior Cloning LSTM. The results demonstrated show the method meeting or exceeding the performance of the compared algorithmic methods. Further, they demonstrate the relationship between the number of fixation transitions made before finding the target and the success rate of the searches. Additional tests were performed, which can be found in the original text [165].

*3.2.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [41].

### 3.3 Active Learning

*3.3.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [98, 118, 121, 133].

*3.3.2 Principles and Definitions.* AL is intended as an AI paradigm by which the learner is able to query an oracle regarding unlabeled data. The goal is to enable a more efficient usage of potentially costly data by allowing the learner to identify a representation of confidence/understanding to determine which items should be prioritized or require input from an oracle [133]. This is described as a characteristic similar to the concept of curiosity in humans, since AL demonstrates a motivation to inspect items with less experience or certainty. Note that this relates to how the system decides what it wants to learn more about, not a true reproduction of human cognition.

However, it has been argued that humans must regulate their priorities regarding curiosity to know when and what to learn [143]. When prompted, the oracle can then provide labels for the queried samples, reducing the need to label larger sized datasets prior to the start of training.

There are numerous ways in which the learner can represent its understanding of the data to determine its confidence level. The representation is used to measure which would be the most desirable query. A possible definition could be the use of entropy to measure the confidence

$$x^*_{ENT} = \underset{x}{\text{argmax}} - \sum_i P(y_i|x;\theta) \log P(y_i|x;\theta) \tag{14}$$

for labels $y_i$, instance $x$, and model $\theta$. This provides an information-theoretic representation of the uncertainty. Specifically, the instance selected for the oracle feedback is the one corresponding to the largest entropy in the probability of assignment to the various labels (as per Equation (14))—that is, the one for which the distribution of probabilities is closer to a uniform distribution (highest uncertainty). Another approach is the instance with the *least confident* labeling:

$$x^*_{LC} = \underset{x}{\text{argmin}}\, P(y^*|x;\theta), \tag{15}$$

where $y^* = \text{argmax}_y P(y|x;\theta)$ is the most likely class labeling. This promotes querying instances that have a best labeling with the least confidence.

The underlying principle in Equations (14) and (15) remains the representation of how well the model can relate instances to the labels based on a self-aware measure of certainty. The utilization of this notion of confidence is what allows for systems to exhibit behaviors akin to curiosity, allowing them to guide the learning process and reduce the reliance on large datasets.

*3.3.3 Applications and Recent Results.* As a demonstration of AL, we discuss the approach outlined in the work of Navidi [99] and Navidi et al. [100], in which a method of Active IL in an RL context is presented. The proposed approach is a divergence from the preceding examples of approaches but still utilizes the general principles of AL. In this case, a model is generated to the oracle's responses for improved performance and to compensate for potential delays between agent action and oracle response. The prediction model enables the agent in its learning process and guides policy learning. Unlike the form of AL described earlier, the oracle feedback provides labels of good or bad behavior, so the agent's policy learning phase will learn to avoid poorly performing actions rather than querying low-confidence items. Hence, the behavior of the oracle encodes a model of the human's preferences with respect to agent behavior. This creates a method combining concepts from IL, AL, and RL. Regarding the underlying algorithm, a combination of SARSA and A3C with human-in-the-loop training was proposed.

To provide feedback from the teacher, agents observe a binary signal indicating a positive or negative teacher response. This is in contrast to other methods, which might provide more complicated or graded feedback by utilizing varying levels of good or bad (e.g., $[-100, -50, 50, 100]$). To compensate for variance in both reaction time and frequency of responses from the teacher (e.g., some might tire of giving feedback), a feedback prediction system is used.

The feedback predictor or manager $FB$ is designed to learn and predict the feedback behavior of the human teacher. $FB$ is provided responses from $\{-1, 1\}$ to signify satisfaction: 1 or dissatisfaction: $-1$. The feedback prediction policy is denoted as

$$FB(O, A) = \psi^T \theta(O, A), \tag{16}$$

where $FB(O, A)$ denotes the teacher feedback policy, $\psi$ are the policy parameters, and $\theta(O, A)$ represents a density function modeling the human response delay.

The feedback is then used to guide the training of the RL policies for the agent. Testing is performed utilizing the SARSA and A3C algorithms. The feedback weights past observations to give

a teacher-based representation of the outcomes. These methods are tested against Open-AI Gym environments such as the Cart-Pole and Mountain-Car scenarios. In both cases, the proposed methods perform strongly in the environments and against the baseline methods.

*3.3.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [41].

## 4 BELIEF AND REASONING APPROACHES

In the following sections, we discuss methods that use representations of belief or world knowledge to guide/assist in agent learning. Relating to the concept of cognitive frugality, we will discuss Meta-reasoning and Meta-learning in Section 4.1. These concepts relate to how skills or knowledge are utilized based on the current context. Further, these concepts also relate to how and when cognitive resources should be applied to determine which system will select the course of action. Compensating for the behavior or preferences of others is another important aspect of modeling and replicating human behavior. The topic **Theory of Mind (ToM)**, which we discuss in Section 4.2, generates a model of the mental states of others to anticipate and collaborate with others. Last, we discuss methods integrating a model of knowledge or world dynamics in Section 4.3. These attempt to replicate human understanding of the world to perform reasoning at a level utilizing these models rather than attempting to learn them indirectly through interactions with the environment.

### 4.1 Meta-reasoning and Meta-learning

*4.1.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [2, 27, 44, 50, 62, 64, 75, 111, 154].

*4.1.2 Principles and Definitions.* Meta-reasoning and Meta-learning consider topics related to a level of abstraction that allows the algorithm to make determinations at two levels. In Meta-reasoning, this is often referred to as "reasoning about reasoning" [27]. Similarly, Meta-learning is often referred to as "learning to learn" [75]. These indicate the core aspect, which is the ability to perform introspection to dictate behavior. More concretely, this often relates to an ability to determine how resources or behavior policies are allocated for reasoning or learning. This implies a notion of cost/effort regarding the task and an ability to identify the most suitable solution/behavior based on context. The notion of allocation for reasoning can also be related to examples in human cognition and the notion of bounded rationality [173], which we discuss in later sections.

*Meta-reasoning.* As noted in the work of Russell and Wefald [126], real agents are limited in capacity with respect to reasoning. Such a limitation manifests in both computational power as well as the time to decide and act. Further, the benefit/utility of an action can deteriorate over the time it takes to deliberate or to execute the action. Consequently, there is an implicit trade-off between the cost of an action (deliberation and execution) and its intrinsic utility. The ability to make judgments, consciously or otherwise, regarding the appropriate balance of these factors is a key aspect of Meta-reasoning. In the context of computer-generated solutions, the common approach is to consider minimal time for an acceptable solution or maximizing outcome at the expense of time to completion. The ability to find an appropriate balance is one of the key motivations for investigating Meta-reasoning topics.

From an algorithmic perspective, the Meta-reasoning problem can be viewed as a method of optimizing the expected utility of performing an action versus the cost to do so. Following Griffiths et al. [50], this can be expressed as

$$\text{VOC}(c, b) = \mathbb{E}_{p(b'|b,c)}[\max_{a'} \mathbb{E}[U(a')|b'] - \max_{a} \mathbb{E}[U(a)|b]] - cost(c), \tag{17}$$

where $b$ is the agent's current belief, $b'$ is the refined belief resulting from executing computation $c$, and $\mathbb{E}[U(a)|b]$ is the expected utility of taking action $a$ under utility $U$ over the distribution of outcomes corresponding to belief $b$. Given the VOC, a rational agent should select the action that maximizes its value (or perform no action if VOC $< 0$). Unfortunately, calculating the VOC is costly, and so methods need to be developed to approximate it or define a similar measure of cost versus utility.

*Meta-learning.* Similar to Meta-reasoning, the case of Meta-learning relies on an ability to perform introspection to understand how to best utilize the resources at hand. In this context, resources are being applied to behavior learning and utilization. As noted in the work of Hospedales et al. [62], Meta-learning involves improving a learning algorithm over multiple learning episodes at two levels. First, in base learning, a learning algorithm learns to solve a task based on the scenario parameters (e.g., image classification). For the second phase, Meta-learning, an outer level algorithm improves the inner learning algorithm. The improvement is made so the resulting learned model improves one of the outer objectives. There can be multiple outer objectives necessitating the algorithm consider which model/algorithm to apply in a given context. Similar to Meta-reasoning, the accrual of information or the time for deliberation can be costly. Consequently, the algorithm requires a method by which it can make a determination regarding the cost/benefit trade-off between the two [44, 50].

For a more formal characterization, we utilize the conventions provided by Vanschoren [150]. Consider the accrual of behaviors for given tasks $t_j \in T$, where $T$ is the set of all known tasks. Given a set of learning algorithms parameterized/configured by $\theta_i \in \Theta$ and evaluation measures $P_{i,j} = P(\theta_i, t_j) \in \mathbf{P}$, a meta-learner $L$ is trained to predict recommended configurations $\Theta^*_{new}$ for a new task $t_{new}$. In this paradigm, it becomes a matter of learning a function $f : \Theta \times T \rightarrow \{\theta^*_k\}$, $k = 1 \ldots K$ that generates configurations $\theta^*_k$. This allows for the creation of a *portfolio* of configurations associated to tasks which they are best suited. This paradigm allows for the generation and selection of configurations based on the experiences of the algorithm. Hence, the learner develops an understanding of what to learn and how to learn it.

*4.1.3 Applications and Recent Results.* Lange and Sprekeler [80], Memory-based Meta-learning (related to the concept of "learning to learn") is investigated. In this scenario, the key concept is an algorithm that can determine when to rely on a learning algorithm for generating behavior versus utilizing a heuristic. This enables agents that develop policies when it is feasible given the time/computational constraints and rely on heuristics otherwise. Regarding learning behavior policies, three features of particular interest are noted relating to the dependence of the Meta-learning algorithm on the meta-RL problem:

- *Ecological uncertainty*: How diverse is the range of tasks the agent could encounter?
- *Task complexity*: How long does it take to learn the optimal strategy for the task at hand?
- *Expected lifetime*: How much time can the agent spend on exploration and exploitation?

Based on the analysis performed in [80], non-adaptive behaviors are optimal in two cases: low variance across tasks in the ensemble, and when time constraints prevent sufficient time for exploration.

For a first scenario, the approach is tested on a two-arm Gaussian bandit task. It is noted that this allows for an analytical solution of optimality. The agent interacts with the environment by performing a series of $T$ arm pulls with rewards represented by a deterministic reward of 0 for the first arm and a Gaussian distribution with variable mean $\mu$ for the second. The mean is sampled $\mu \sim \mathcal{N}(-1, \sigma_p^2)$ (where $\sigma_p$ specifies the scale of ecological uncertainty) at the beginning of each episode, and is then used to define the reward function $r \sim \mathcal{N}(\mu, \sigma_l)$. It is noted that this variability controls how many arm pulls are needed to estimate the mean reward and $\sigma_l$ controls how quickly

the agent can learn the policy as it controls the consistency of the observed rewards. The scale of $\sigma_p$ clearly determines how much uncertainty should be expected for the agent regarding the mean reward observed from the stochastic arm, which consequently scales the difficulty of estimating the expected utility of this arm. For $\sigma_l$, this determines how difficult it will be to estimate the mean based on observed rewards.

Given this definition, the optimal solution is determined analytically based on the problem characteristics. Since the agent should perform $n$ pulls for exploration before exploiting its knowledge, the goal is to identify the optimal number of trials $n^*$ before concluding the exploration. The solution is as follows:

$$\begin{aligned} n^* &= \underset{n}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=1}^{T} r_t | n, T, \sigma_l, \sigma_p\right] \\ &= \underset{n}{\operatorname{argmax}} \left[-n + \mathbb{E}_{\mu,r}\left[(T-n) \times \mu \times p(\hat{\mu} > 0)\right]\right], \end{aligned} \tag{18}$$

where $\hat{\mu}$ is the estimated mean reward of the second arm after $n$ exploration trials.

Based on experiments, two distinct types of behavior were noted. In the first, learning via exploration is the optimal behavior. In the second, the agent should not learn and instead exploit its knowledge. The value in stopping learning is attributed to two aspects: (1) small ecological uncertainty can make it very unlikely the first arm is better, and (2) too high of variance causes the lifetime to be too short for valid learning. In other words, uncertainty and resource limitations can cause a shift in the prioritization of learning and knowledge exploitation. As a result, that this leads to two consequences. First, the values of $\sigma_l$ and $\sigma_p$ create a threshold between learning and not learning of behaviors. This determines whether learning is suitable or if the agent should instead rely on exploitation of the existing knowledge. Second, the optimal strategy is dependent on the time allotted for learning. The amount of knowledge an agent can attain is strictly dependent on the variance parameters and the time they have to learn the dynamics of the arms. The results indicate the relationship between complexity and uncertainty with respect to the expected number of trials. There is a clear delineation between the two behaviors indicated in the results.

Given the preceding specification, agents trained on the bandit scenario, were generated, which were subsequently compared to the theoretically optimal solution. The outcome indicates the Meta-learning paradigm creates agents matching the analytical solution. Given this structure, [80] notes that such an agent demonstrates the dual components that handle learning a behavior policy and utilizing a hard-coded choice that selects the deterministic arm (optimal choice in expectation).

To test the proposed approach in a more general case, an LSTM-based actor-critic agent is trained and tested in an ensemble of gridworld tasks. This scenario allowed investigation of the impact of lifetime on the exploration strategies of the agents generated. Intuitively, in the case of a long lifetime, the agent has more opportunity to search for higher-valued goals where shorter lifetimes would necessitate greedier identification of goals. This was tested by placing goals of increasing value at farther distances from the start state.

As indicated by the results, the trained agent demonstrates a learned preference for goals based on the proximity and lifetime $T$ provided. The results indicate the agent is able to learn a priority for farther goals when there is more time to discover and return to the higher-valued goal. Similarly, the agent learns to focus on goals closer to the start state when there is a higher restriction on $T$ or when the agent should prioritize the goal states based on time remaining. The agent exploits its policy to find the highest value goal while there is time, then switches to the lower value state when there is only time for these path lengths. This indicates the agent is able to exploit different experiential knowledge while also identifying when to switch between exploration or learned

behaviors. Based on the two scenarios, the results demonstrate the method having the means to learn how to switch between learning and exploiting learned models. This demonstrates "learning to learn" and an ability to minimize the cost of behavior.

*4.1.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [41].

### 4.2 Theory of Mind

*4.2.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [17, 48].

*4.2.2 Principles and Definitions.* ToM is what gives humans the capacity for reasoning about the mental states such as beliefs and desires for other agents in their environment [13, 40, 115, 155]. In most, if not all, aspects of daily life, humans rely on and utilize their ability to estimate the mental state of others. One can imagine numerous scenarios they might encounter daily where they interact with someone while relying on this type of reasoning. Something as simple as trying to anticipate on which side to pass another pedestrian requires this technique. It is also easy to imagine a more complex scenario, such as poker. Players must reason about both the hand likelihoods as well as the mental state of their opponents to ensure a higher chance of success. In poker, players can utilize methods of deception in an attempt to disguise their true mental state and gain an advantage. As a result, players need to reason about the mental states of their opponents to guide their playing strategy.

ToM relies on multiple aspects of perception. A person will utilize the non-verbal cues, contextual clues, and other stimuli to form a picture of the world from the perspective of another. ToM is what allows you to imagine yourself as someone else to model the likely next steps. Such a skill allows humans to perform better both as individuals and as part of a team. As with the walking and poker examples, one could also imagine numerous scenarios where people work together to accomplish a goal. People working together will of course utilize verbal and other forms of communication, but there is also a significant reliance on each member's ability to use a deeper understanding of the observed behaviors of their collaborators. Without this ability, we could expect that every aspect of these interactions would require explicit and comprehensive communication regarding all aspects of the interaction.

The representation of others is not necessarily exact, but can instead utilize an approximate and higher-level model [115]. Further, it can be argued that humans bias their models of others based on their own perspective. It is natural to expect a person's past experiences to impact their model of another person's perspective. Given a system to model the perspective of others, the natural extension is generating recursive relationships between them [155]. For instance, modeling the mental state of others can be extended to model the other person's model of one's self. This means a person can generate a model of how they are perceived from another person's perspective. As noted by Freire et al. [40], several notable approaches exist to address the problem of ToM. These include methods relying on RL, neural networks, policy reconstruction, and so forth.

*4.2.3 Applications and Recent Results.* One aspect of ToM is predicting the behavior of others so we can act accordingly. The work presented in Wang et al. [156] demonstrates a use case in which an RL-based learner predicts movements of another agent to support rendezvous in a multi-agent environment. This is accomplished by generating a model for motion prediction that supports a Hierarchical Predictive Planning (HPP) module.

Formally, the problem is defined as a **Decentralized Partially Observable Markov Decision Process (Dec-POMDP)** $\mathcal{M} = \langle n, \mathcal{S}, O, \mathcal{A}_{1,\ldots,n}, T, \mathcal{R}, \gamma \rangle$, where $n$ is the number of agents, $\mathcal{S}$

denotes the set of states, and $O = [O_1, \ldots, O_n]$ is the joint observation space ($O_i$ is agent $i$'s observations). The joint action space $\mathcal{A}_{1,\ldots,n}$ is the Cartesian product of the agent action spaces $\mathcal{A}_i$ for agents $i \in \{1, \ldots, n\}$, which define the actions available to the agents. The transitions are defined by the function $T : \mathcal{S} \times \mathcal{A}_{1,\ldots,n} \times \mathcal{S} \rightarrow [0, 1]$, which models the probability of transitioning between states given a joint action $a_{1,\ldots,n}$. $\mathcal{R}$ is the reward function and maps states and actions to a reward $r \in \mathbb{R}$. As defined in RL, $\gamma$ is the discount factor for the learning process. The assumption in Dec-POMDPs is that agents receive noisy observations of the environment, so the true state $s_i$ is unknown and is instead represented by the observation $O_i$. Consequently, the behavior relies on the estimated current state to identify a desirable action.

The observations provided based on this method represent the beliefs regarding the position and observations of each agent. This enables generating predictive models that can attempt to reduce the error in predictions of future state for one's self or others. The future predictions allow an estimation of likely positions and observations for future timesteps, and the error of the prediction can guide the updates to the learning method in the form of a reward. In this case, the reward is the negative sum of the prediction error, which motivates the agents to improve their ability to predict the movements of others. The learned predictive model enables estimation of likely rendezvous locations in this scenario.

Agents are provided observations $O_i = [\mathbf{p}_i, \mathbf{p}_{-i}, \mathbf{o}, \mathbf{g}]$, where $\mathbf{p}_i$ and $\mathbf{p}_{-i}$ refer to the agent positions, $\mathbf{o}$ are the sensor observations, and $\mathbf{g}$ is the agent's goal. The agents learn a predictive model of motion via a self-supervision algorithm. The systems are tasked with learning two models: *self-prediction*($\mathbf{f}_i$) and *other-prediction*($\mathbf{f}_{-i}$). In this context, $\mathbf{f}_i$ and $\mathbf{f}_{-i}$ refer to self and other dynamics models, respectively. Note that these models are the key point where ToM is used in this example as they enable modeling and prediction of others. These models are learned to generate a self-prediction of position $\Delta\mathbf{p}_i^{t+1}$ and observation $\Delta\mathbf{o}_i^{t+1}$, using a time window of past position that covers the previous $h$ steps:

$$\left(\Delta\mathbf{p}_i^{t+1}, \Delta\mathbf{o}_i^{t+1}\right) = \mathbf{f}_i\left(\mathbf{p}_i^{t-h:t}, \mathbf{o}_i^{t-h:t}, \mathbf{g}\right). \tag{19}$$

Similarly, a model for other-prediction (i.e., other agent prediction) is defined as

$$\left(\Delta\mathbf{p}_{-i}^{t+1}, \Delta\mathbf{o}_i^{t+1}\right) = \mathbf{f}_{-i}\left(\mathbf{p}_{-i}^{t-h:t}, \mathbf{o}_i^{t-h:t}, \mathbf{g}\right). \tag{20}$$

It is noted that the models do not depend on actions, but instead utilize observations of positions and pose conditioned on goals, which prevents needing to know the action space of others.

Given the predictive models $\mathbf{f}_i$ and $\mathbf{f}_{-i}$, a decentralized policy $\Pi_i$ is generated via the Cross-Entropy Method (CEM) to convert goal evaluations into belief updates over potential rendezvous points. The demonstrated intuition behind this approach is that each agent is simulating a centralized agent that fixes the goal of all agents, which are used to precondition the motion predicted by $\mathbf{f}_i$ and $\mathbf{f}_{-i}$. These predictions are performed in a rollout of $T$ timesteps into the future to predict the pose of agents. Based on the rollouts, the predicted goals are scored according to

$$\mathcal{R}(\mathbf{p}_{1,\ldots,n}) = \begin{cases} 0, & |\mathbf{p}_j - \mathbf{p}_\mu| < d,\ \forall j \in 1, \ldots, n \\ \sum_{j,k\neq j} -|\mathbf{p}_k - \mathbf{p}_j|, & otherwise \end{cases} \tag{21}$$

where $\mathbf{p}_\mu = \frac{1}{n}\sum_{k\in 1,\ldots,n} \mathbf{p}_k$ and $d$ is a precision parameter. This emphasizes accuracy of predicted future states and smaller distances between agents at rendezvous points. Based on the predicted goal values, goal states are sampled via a normal distribution and the estimation process continues to favor goals that bring the agents closer together. This process generates the policy $\Pi_i$, which is used to complete the rendezvous without centralized control. With the preceding approach, a ToM

method is demonstrated utilizing the first level of ToM reasoning to predict likely agent behavior based on past observations and current circumstances.

The algorithm's capabilities are demonstrated in simulated and real-world environments. Simulated environments range in complexity starting with no obstacles and transition to environments with multiple obstacles. Similarly, the physical environments vary in obstacle complexity, type, and layout. The approach was tested against learned, decentralized, planning-based, and centralized baselines. For the learned baseline, the popular MADDPG algorithm is used. Planning was performed using the RRT planning system. For the centralized system, the midpoint, other agent's position, and random point are used. As is clear from the results, the proposed approach performs strongly compared against the baselines. This demonstrates the agents are able to perform the rendezvous without the need for centralized control. In the case of the real-world environments, we can see a similar effectiveness of the proposed approach. The results indicate that the approach is able to translate from simulated environments into the real world.

*4.2.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [41].

### 4.3 Simulating Human Knowledge of World for Learners

*4.3.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to the work of Fuchs et al. [148].

*4.3.2 Principles and Definitions.* The topics in this section do not represent a comprehensive list, but instead serve to illustrate a type of learning that relies on a level of world comprehension to accomplish the learning task. In this case, learning agents are provided models allowing them to understand a feature of the environment instead of requiring the agent to learn these features. For example, humans demonstrate a capacity for modeling the fundamental characteristics of an environment such as physics and compositional structure (e.g., *Solidity* [148]). With such an understanding, the components of a scenario can be considered when learning or utilizing a skill. Such a behavior is demonstrated in multiple aspects of daily life. Something as simple as understanding that gravity allows a person to pour water into a glass is something we take for granted. Comprehension of these aspects of the world enables a rich set of behavior. For artificial systems, such a comprehension is often not provided or demonstrated. To overcome this, researchers have investigated methods in which features such as physical, compositional, and hierarchical are provided as a model for simulation or planning.

*Physics Models.* Following the assumption that humans possess an internal model or understanding of aspects of physics (e.g., pushing an object over the edge of a table causes it to fall), digital systems can be generated to replicate these behaviors for modeling and simulation. Engines such as Unity or MuJoCo [146] demonstrate motion of objects in physical environments and can provide realism to object motion for observations made by artificial agents. These models of system behavior provide an agent with an internal estimator for world dynamics to support prediction and planning [6, 107, 108]. This lets a learner generate predicted future states for reasoning and planning.

*Planners.* Similar to a physics engine, the components and dynamics of a system can be modeled by describing the objects in the environment and how they can interact, be utilized, or be modified. This allows for constraints specified for states that must be met so an action is available for execution. The combination of the environment composition and the available actions allows for planning. Planning is utilized to convert the system state from the initial configuration to a goal state through a sequence of actions. A commonly used method for representation is via **Planning Domain Definition Language (PDDL)** [3]. An environment specified in PDDL can then

be analyzed by a solution system to identify a desirable sequence of actions for the specified goal. These solutions are generated to utilize the world model and the defined constraints. This enables identification of viable plans without the need for a learning process such as RL. However, policy generation via RL with planning-based trajectories can be performed [171].

*4.3.3 Applications and Recent Results.* To demonstrate the use of a modeling system to enable deeper understanding, we will discuss the work of Zhi-Xuan et al. [171], which integrates PDDL into the model for planning based on different likely goals or trajectories. The inclusion of a planning system was intended to account for suboptimal or failed plans and incorporate them into estimates of future outcomes. Further, this approach was desired as a method accounting for the difficulty of the planning phase itself. In [171], it's noted how many methods attempting to estimate goals fail to consider the difficulty of the planning portion of the process. Additionally, how most methods require the assumption of optimal behavior or Boltzmann-rational action noise. The assumption of optimality for all goals would therefore require computation of all goals in advance, which is intractable. In the proposed approach, the integration of a planning system is used to represent a boundedly rational agent, where the bounds provide a resource limitation on planning and plan execution. This model provides a mechanism for Bayesian inference of plans/goals, even those with suboptimal solutions, requiring backtracking, or irreversible failure. The limitation forces a constraint on the time or resources available to make a decision, which forces the agent at times to generate only a partial plan up to the level afforded by the constraints.

As noted earlier, the proposed approach represents the states, observations, and goals using PDDL and a variant supporting stochastic transitions named *Probabilistic PDDL*. This is accomplished by representing states and goals via predicate-based facts, relations, and numeric expressions in PDDL format. This allows for modeling of world state and actions, which are available when the provided preconditions are satisfied (current state, tool availability, etc.). The combination of predicates (e.g., relations like "on" or "at") and fluents (e.g., fuel level) allows the planning system to identify system state and available resources to generate a chain of actions and outcomes leading to a goal state. The actions result in a change in predicates and fluents, which signify the transition between world and system states. For a simple example, consider stacking blocks. A block that is covered by another would not be available for stacking, so the predicate's constraint would not be met for this block. However, uncovered or top-most blocks would be available, so one from this set would be available for selection for movement. With this representation, the observer has a prior over goals $P(g)$ specified via a probabilistic program over PDDL goal specifications.

To represent bounded rationality, a budget is defined via:

$$\eta \sim \text{NEGATIVE-BINOMIAL}(r, q), \tag{22}$$

where $r$ denotes the maximum failure count and $q$ denotes the continuation probability. Therefore, $\eta$ sets an upper bound on the solution search. If the bound is reached, the agent executes a partial plan leading to the most suitable state reachable from the found plans. This allows the agent to find the best plan they can given the limited resources available. Given the proposed approach, it is noted this model supports any planner capable of producing partial plans. For the given scenario, agents are operating in a gridworld environment and utilize a variant of the $A^*$ algorithm [54] that makes search stochastic.

To test the proposed **Sequential Inverse Plan Search (SIPS)** approach, several environments were utilized with goal set $G$ and state space $S$:

- *Taxi* ($|G| = 3$, $|S| = 125$): A taxi has to transport a passenger from one location to another in a gridworld.

- *Doors, keys, and Gems* ($|G| = 3, |S| \sim 105$): An agent must navigate a maze with doors, keys, and gems.
- *Block words* ($|G| = 5, |S| \sim 105$): Goals correspond to block towers that spell one of a set of five English words.
- *Intrusion detection* ($|G| = 20, |S| \sim 1030$): An agent might perform a variety of attacks on a set of servers (20 possible goals corresponding to a set of attacks on up to 10 servers).

The approach is tested against a **Bayesian Inverse Reinforcement Learning (BIRL)** model generated by value iteration. This allows for comparison against an approach based on RL solutions. As a demonstration of effectiveness, the performance of SIPS and BIRL were compared to human performance in goal prediction. The results indicate a strong match between the SIPS model and the demonstrated human performance. These results were generated by collecting human goal inferences on 10 trajectories with six suboptimal or failed, for $N = 8$ subjects. Human inferences were collected every six timesteps. For evaluation of accuracy and speed, tests included a dataset of optimal and non-optimal trajectories. The optimal trajectories were obtained via $A^*$, and the non-optimal trajectories were generated using a replanning agent model with $r = 2, q = 0.95, \gamma = 0.1$. Inference is performed on these datasets with a uniform prior over goals. Based on the tests, results indicate good performance with 10 particles per goal without the use of rejuvenation moves. In both test cases, the proposed method demonstrates strong performance. Additionally, the performance versus computational cost is quite strong in the majority of cases in comparison with the baseline BIRL.

*4.3.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [41].

## 5 BOUNDED RATIONALITY AND COGNITIVE LIMITATIONS

In the following sections, we will discuss methods inspired by the cognitive limitations and characteristics demonstrated in human reasoning. The approaches discussed in Section 5.1 demonstrate how humans utilize heuristics to enable fast and frugal reasoning as well as more deliberative systems. This combination of systems enables more efficient use of cognitive resources and has been demonstrated in numerous studies of human reasoning. Related, we see the use of systems designed to replicate the cognitive and neurological performance (not necessarily the physical structure) seen in humans (see Section 5.2). These systems are designed to mimic human performance on tasks by replicating how humans use knowledge and memories to make decisions and perform tasks. Next, in Section 5.3, we discuss techniques inspired by how humans attend to stimuli and make associations between observed values. This can relate to human vision (i.e., foveation) or how humans identify correlations between different items in the same context (e.g., the word *book* in a sentence would increase the relevance of the word *library* in the same sentence). Last, we will outline concepts relating to bounded rationality in the context of Game Theory in Section 5.4.

### 5.1 Cognitive Heuristics

*5.1.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to the work of Booch et al. [19].

*5.1.2 Principles and Definitions.* Bounded rationality describes the notion of humans making rational choices under the constraints ascribed to cognitive limitations of the decision maker [12, 124, 135]. These constraints are a reflection of the assumed limitations or deficiencies in a human's computational abilities/capacity and knowledge. Similarly, these constraints can be viewed as respecting a notion of cognitive or computational cost. As a reflection of these limitations, there is an assumption that humans perform decision making in a manner allowing

them to find a reasonable approximation of the optimal solution while reducing overall cost or time. Reasonable could be viewed as a "good enough" or *satisficing* solution. Finding such an acceptable result is supported by shortcut techniques (i.e., heuristics) that allow a person to approximate the collection of alternative solutions by finding a set of satisfactory alternatives.

An additional aspect of bounded rationality is the assumption that humans will sacrifice exactness/optimality of a solution for the sake of efficiency. In this context, efficiency can refer to time required to a solution, cognitive resources needed, and so forth. In general, this alludes to a sense of frugality when it comes to the cognitive resources a human is willing to dedicate to a decision process. In most circumstances, this frugality does not cause issues and allows for sufficient stimulus processing. However, there are examples of where this may cause a significant omission of perception. For example, humans can be tasked with observing a scene and then asked questions at the end of the observation pertaining to specific content. Often, there can be items or people hidden in plain sight due to the observer being distracted by more attention-grabbing stimuli, which demonstrates possible errors in the use of cognitive heuristics by the human brain [153].

*Dual-System Reasoning*. Relating to bounded rationality, there has been extensive interest in what is referred to as dual-system reasoning or dual-process theory, now extending to topics in AI [19]. The argument is that humans utilize two systems of reasoning based on the context of the problem and the limitations of their cognitive systems. The assumption is that the two levels handle problems at different speeds, fidelity, cognitive cost, and so forth. These distinctions are based on the belief that humans tend to utilize a lower-cost reasoning system when the penalty for a suboptimal solution is minor or when the time or cognitive burden of reaching a higher accuracy solution is too great (see other works [92, 111] for more on cognitive cost). For simplicity, we follow a common convention and refer to them as System-1 and System-2. For an example of System-1 reasoning, catching a falling object typically does not leave sufficient time for deeper reasoning, so we rely on instinctive movements made quickly by System-1. However, System-2 can support deeper reasoning, longer time to a decision, and so forth.

System-1 is commonly assumed to be based on approximations generated via heuristics. These heuristics provide shortcuts to reasonably accurate solutions. For example, it has been argued that humans utilize what is known as the availability heuristic, which selects a solution based on the strongest association between the current situation and memories of potentially similar instances. This means humans will tend to place higher weight on memories more closely aligned with the current observations. As such, they will be biased toward solutions with higher likelihood of recall. This allows humans to use similar past experience to simplify the decision-making process when using System-1. If the problem is too complex for System-1, then System-2 needs to be utilized for deeper inspection and the possibility to use/combine multiple underlying processes.

The use of these systems can lead to biases in reasoning and potential incorrect assessments. A simple example is the Gambler's Fallacy, which demonstrates how humans tend to believe that a sequence of flips from a fair coin should be self-correcting [69, 147]. In other words, when the coin is flipped multiple times, a sequence of identical outcomes is considered less and less likely as the length of the sequence grows. This misconception leads the person to feeling that the alternate outcome should be more likely in the next instance. As is apparent from the independence of the samples, this is in fact incorrect reasoning. Probability theory dictates that the outcomes of each toss should have no effect on the next toss.

Humans demonstrate additional forms of heuristic-based and biased reasoning. The following are some additional examples [38]:

- *Satisficing*: Use a sufficient option rather than the optimal one.
- *Affect*: Make a decision based on intuition or "gut feeling."

- *Availability*: Estimate the likelihood of a future event based on the strength of recall for similar past occurrences.
- *Representation*: Assume X is the same as Y when you notice X is similar to Y in some way(s).
- *All or nothing*: Simplify decisions by treating remote probabilities as if they were not even possibilities.

These heuristics/biases demonstrate systems that can generate correct solutions in many cases, but can also lead to misconceptions or ignored information. As a result, it stands to reason that methods hoping to model the behavior and reasoning of humans will need to take these potential inaccuracies into account.

*5.1.3 Applications and Recent Results.* Lieder et al. [85], the proposed approach demonstrates an application of the dual-process concept of human reasoning and use of heuristics in an RL paradigm demonstrating multi-alternative risky choice in the MouseLab scenario (widely used to study decision strategies). In this scenario, the approach demonstrates the emergence of two known heuristics: **Take-the-Best (TTB)** (chooses alternative favored by the most predictive attribute and ignores others) and random choice. It is noted that these are resource-rational strategies for low-stakes decisions with high and low dispersion of their outcome probabilities, respectively. Further, how the TTB heuristic is commonly used by humans when under time pressure and one outcome is much more likely than others. Similarly, [85] demonstrates: how humans tend to accept random selection when stakes are low in low-dispersion cases. The bounded optimal decision process is represented as a meta-level MDP by considering the cost of computing a solution that impacts the utility of a decision or action. The actions are treated as costly computations, necessitating the ability to make decisions with efficiency in mind. This need for efficiency follows those seen in the justification of the representation and use of heuristics.

An augmentation to the MDP considered in this research is the meta-level MDP. In this case, actions for the meta-level MDP are cognitive operations $C$ performed in belief states $b_t \in \mathcal{B}$. Additionally, the meta-level MDP has a transition function $T_{meta}$ and reward function $r_{meta} \in R_{meta}$. The operations in $C$ include an operator $\perp$ that terminates deliberation and subsequently translates the current belief into an action. The determination to end deliberation and select an action can be seen as a representation of how humans select System-1 or System-2 reasoning, which then results in an outcome from the selected system. The reward $r_{meta}$ combines the cognitive cost $c \in C$ with the expected immediate reward the agent expects to receive once deliberation terminates and an action is taken. In the case of a computation, the reward is defined as $r_{meta}(b_t, c) = -\text{cost}(c)$ for $c \in C$; otherwise, $r_{meta}(b_t, \perp) = \text{argmax}_a\, b_t^{(\mu)}(a)$, where $b_t^{(\mu)}(a)$ is the expected reward of action $a$ according to belief $b_t$.

The MouseLab scenario provides a testbed in which agents can improve the likelihood of successful decisions by performing additional information acquisitions. Although the acquisition improves the decision, it also incurs a cost. Therefore, the agent should minimize the occurrence of cognitive costs while maximizing the subsequent game outcome. This promotes a trade-off of decision quality and decision time, mimicking the similar processes witnessed in human cognition. As noted in the results, the proposed method rediscovered TTB, WADD (Weight-Additive Strategy), and then random choice strategy. The additional strategy, WADD, is performed by computing the expected values of all gambles using all possible payoffs.

There were three noted outcomes regarding the predictions and the pattern that justify use of heuristics and match the observation of study participants (200 participants on Amazon Mechanical Turk). First, the model predicted fast-and-frugal heuristics should be prioritized/utilized more frequently in high-dispersion trials (high dispersion means an outcome significantly more likely than the others, and fast-and-frugal heuristics ignore all outcomes except the most probable).

Second, the model indicates the utility of simple heuristics, primarily when the stakes are low. Third, the model indicates the benefit of increased time and effort for high-stake scenarios to receive the highest possible payoff.

*5.1.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [42].

### 5.2 Cognitively/Biologically Plausible Representations

*5.2.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to the work of Kotseruba and Tsotsos [78].

*5.2.2 Principles and Definitions.* With the goal of achieving a general AI (i.e., reaching human-level intelligence [86]), there have been numerous approaches inspired by the cognitive mechanisms enabling the intelligence observed in humans. Russell and Norvig [125], several approaches are noted regarding how reaching human-level general intelligence might be possible. One noted method relates to the design and justification of cognitive architectures. For cognitive architectures, the goal is not always to achieve a perfect analog of the human brain and its neurological function; instead, a common goal is to generate a system capable of demonstrating the same kinds of abilities and deficiencies seen in human cognition, reasoning, intuition, and so forth (e.g., perception, memory, attention) [34, 46, 72, 78, 145]. Under these circumstances, the goal is often the creation of a model of behavior that fits the cognitive/neurological dynamics of the human brain [7, 11, 78, 123, 137, 140, 149, 159].

As noted in Section 5.1.2, it is generally accepted that humans reason with systems operating at different levels of fidelity. Humans can make faster and cognitively frugal decisions or utilize slower and more cognitively burdensome resources. As such, research has been dedicated to the creation of systems demonstrating these characteristics (and beyond) [78]. These systems demonstrate an ability to learn behavior as we have seen in previous sections (e.g., RL), but the distinction in this case is the emphasis on replicating the cognitive performance of humans. This distinction motivated us to place a higher emphasis on the cognitive and biologically plausible mechanisms of this portion of the article. Further, there have been studies that show these representations can provide the best-performing (and likely best-fitting) approximations to human cognitive performance [138]. For example, cognitive architectures utilize memory systems to replicate how humans retain information and utilize that information when making decisions. As a result, we see this section as more suitable in a cognitive limitations and biases context.

*World Representation: Symbolic, Emergent, and Hybrid.* To support reasoning and behavior, the system needs a method for representing the world. For cognitive architectures, there are three main categories for the underlying representations: symbolic, emergent, and hybrid. As the name would suggest, symbolic systems use symbols to represent concepts or knowledge. Given the symbols, the system can manipulate them using a given set of instructions provided through if-then rules or similar means. As can be expected, a symbolic representation allows for accurate planning and reasoning, but the potential downside being this approach is brittle and does not adapt to changes in the environment. Emergent systems operate similar to what is seen in Artificial Neural Network (ANN) systems. Information is processed by the system, and associations are made through a learning process. This of course increases the system's responsiveness to changes in an environment, but can reduce the transparency or the ease of interpreting the system's behavior. To utilize the advantages of both systems, with the hope of overcoming the shortcomings, there are hybrid systems combining the symbolic and emergent approaches.

*Learning Methods.* Cognitive architecture can perform learning in several ways, including declarative, procedural, associative, and so forth [78]. In the declarative case, the system is provided a collection of facts about the world as well as relationships between them. For instance, many systems, such as **Adaptive Character of Thought–Rational (ACT-R)**, SAL, CHREST, or CLARION, utilize chunking mechanisms to declare new knowledge items. For the procedural case, the system learns skills gradually through repetition, which can be accomplished through the accumulation of examples of successful execution of a task or problem. More closely aligned to RL, the associative case is based on observations of rewards or punishments.

*Memory.* Architectures can be supported by different memory mechanisms depending on the type of capabilities being replicated. When performing a task, the memory utilized to temporarily store information related to the task at hand is referred to as *working memory* [8]. This memory is updated rapidly as the state of the world changes and actions are taken. Further, there is commonly an assumption regarding the capacity limitations of working memory for humans. In addition to working memory, other systems provided a means to accomplish long-term memory storage. This can support storage of procedural memory to define basic skills or behavior or declarative memory for knowledge. This allows for the storage of innate skills as well as accumulated knowledge. Additionally, some systems are defined with a hybridization of long- and short-term memory, referred to as *global memory*. This results in all knowledge and memories being represented by the same system.

*5.2.3 Applications and Recent Results.* The preceding characteristics are broad aspects covering different approaches for cognitive architectures. For a specific example, we present a recent result based on the ACT-R architecture. The approach in [28] presents a cybersecurity game designed to demonstrate cognitive biases of cyberattackers. This displays how humans are susceptible to fallacies in reasoning which result in suboptimal and biased behavior. The approach demonstrates how the models used replicate the biases motivating human behavior patterns in system selection and the choice of whether to abandon a system and forfeit the previous effort on the current system.

The proposed approach used an **Instance-Based Learning (IBL)** model using the ACT-R architecture. ACT-R is a theory of cognition that models how humans recall "chunks" of information from memory and problems by splitting them into sub-goals [159]. Knowledge is applied from working memory as needed to find a pattern of behavior meeting the goal. This model utilizes techniques designed to mimic human memory retrieval, pattern matching, and decision making. IBL uses ACT-R's blending mechanism, interpolating across past experiences to estimate an outcome. The interpolation is weighted by the contextual similarity between the present observation or instance and the past experiences. This provides an estimated expected outcome based on the *consensus value* $V$ that minimizes the dissimilarity (measured by $Sim$) from the values contained in instance $i$ defined as

$$\underset{V}{\operatorname{argmin}} \sum_{i} P_i \times (1 - Sim(V, V_i))^2, \tag{23}$$

where $i$ refers to an instance stored as a memory chunk representing a past state-action-outcome observation and $P_i$ refers to the retrieval probability (based on IBL-based measures). In the case where $Sim$ is interpreted as the error, then Equation (23) generates a least-squared error method [81]. In other words, this finds an estimated value $V$ that best fits the past observations weighted by their strength of recall. These estimated values are used to make a determination regarding which action/production should be executed. The measure or threshold that determines whether an action is available for execution limits the set of possible actions further. This means that the value is based on a representation that considers how strongly a memory is remembered, how

similar the memory is to the current context, and the value observed by the choice made in that past observation. The strength of a memory represented by the retrieval probability utilizes a Boltzmann softmax equation

$$P_i = \frac{e^{A_i/\tau}}{\sum_j e^{A_j/\tau}}, \tag{24}$$

where $\tau$ defines the temperature parameter, which scales probabilities defined by the activation function. The activation function provides a measure of how strongly a memory is remembered and associated with the current context. This strength is based on elapsed time since the observation was made. The activation for a chunk or instance $i$ is defined as

$$A_i = \ln \sum_{j=1}^{n} t_j^{-d} + MP \times \sum_k Sim(v_k, c_k) + \epsilon_i, \tag{25}$$

where $t_j$ refers to the elapsed time since the $j^{th}$ occurrence of instance $i$, $d$ is the decay rate (commonly set to 0.5), $c_k$ refer to the context elements, $v_k$ refer to the instance in memory, and $MP$ is the mismatch penalty (in this case, set to the default of 1.0). The first term in Equation (25) provides the measure of strength based on the time elapsed, and the second term is another similarity term similar to what is seen in Equation (23). $MP$ is a weight term parameter that scales the similarity scores in the sum, and the last term, $\epsilon_i$, is a variance parameter providing stochasticity in the activation function. Similar to Equation (23), the $Sim$ measure ensures the memories considered are a suitable match to the current context to prevent consideration of too dissimilar of instances. In more general terms, the preceding equations define a method for determining which memories are considered, how strongly they impact the estimate based on past observations, and how the resulting behavior occurs based on this historically weighted knowledge.

The agents are trained to perform the cyberattacker role. As such, the agents are provided observation instances that include the probability of a system being monitored, the reward for successful infiltration of a system, the penalty for detection, and a warning signal denoting whether a system is being monitored. The model is then primed with seven instances: five simulating a practice round and two representing knowledge of occurrences (absent and success, absent and uncertainty). This provides the system with an initial set of experience to allow for initialization of learning behavior without relying on random decisions. The model then is trained for four rounds of 25 trials.

The model was tested in comparison to human performance. Human players were studied to generate a baseline of behavior and identify any demonstrated biases in outcomes. Based on the experiments, the results show the human players demonstrating preferences or likelihoods of attack for different systems. They also demonstrated the cognitive systems performing equivalent preferences/probabilities.

*5.2.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [42].

### 5.3 Attention

*5.3.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [26, 48, 102, 103].

*5.3.2 Principles and Definitions.* Humans and other living beings do not process all the available perceptual information available to them. Instead, they utilize a cognitive and behavioral system that allows them to reduce the complexity of perception through an objective or subjective selectiveness with respect to information. This selectivity or bias is referred to as attention [26]. A basic

interpretation justifying the biological need for attention would be the fact that our environment provides more stimuli than we can reasonably process fast enough. In this case, "fast enough" is with respect to the actions or behaviors necessary for survival.

When facing a critical situation, timeliness can be crucial; otherwise, an overload of stimuli could cause a costly delay (e.g., moving out of the path of an oncoming vehicle). As such, our brains allow us to reduce, or even ignore, information perceived to reduce the cognitive burden. Further, attention allows us to prioritize the information and assign more or less significance based on learned/perceived importance. In the oncoming vehicle example, it is likely not important to note the color of a building in the distance while estimating the speed and trajectory of the vehicle.

From a computational perspective, attention initially was studied primarily in the context of vision [26] where images were studied under the task of identifying salient regions. Artificial systems were developed to generate maps that would filter the input for processing. With the growing popularity of Deep Learning, attention techniques were transitioned to neural network paradigms. Attention is used to modify the flow or processing of information in the network(s). The task of learning attention allows the systems to learn how to ignore stimuli, similar to the natural analogs mentioned. This allows systems to contextually alter the significance of information to better suit the underlying task. The attention mechanisms utilized in Deep Learning can be categorized as follows.

*Soft Attention.* Soft attention uses softmax functions to weight the input elements with a weight value in $(0, 1)$. This allows the system to learn and utilize an interdependence between different input parameters. Being based on softmax functions, soft attention provides a differentiable mechanism for attention. The soft attention scales the relative intensity of the input parameters.

*Hard Attention.* Hard attention, as the name suggests, is the complement of soft attention. It utilizes weights in $\{0, 1\}$ to generate a mask to signify whether information is used or entirely ignored. As a result, the hard attention mechanism is non-differentiable. This necessitates a learning process for determining where to assign the weight values. In this case, there is a distinct exclusion of regions of the input domain while the remainder is observed at normal scale.

*Self-Attention.* In self-attention, the system is learning an interdependence between sequential input elements. This allows a system to identify and utilize a notion of relation between items in the same input sequence. As a result, self-attention can be useful in understanding deeper relationships between items in the input rather than a holistic view of the input. For example, Vaswani et al. [151] introduce the transformer network, which performs self-attention using representations of queries $Q$, keys $K$, and values $V$:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{26}$$

where $d_k$ refers to the dimension. In this case, the attention mechanism learns associations between different components of the input and their corresponding strengths via a representation as keys $K$ and queries $Q$. This allows the model to learn a relationship between the current task, the input data, and the current query. The associations are learned as weight matrices, which scale the input values and give the weighted strength of association. This generates a weighted association between elements of the input and forms a compatibility measure of the values.

*5.3.3 Applications and Recent Results.* The approach presented in [105] demonstrates the use of attention with an RL agent to control simulated traffic lanes. The motivation for the use of attention provided references the fact that traditional systems would require retraining for a new lane configuration. The use of attention allows for more flexible representations and can handle

different numbers of roads/lanes. The proposed algorithm is tested against several baselines and demonstrates strong performance in traffic regulation.

For a road with intersection $m \in \mathcal{M}$, define the *traffic characteristics* $s_l^t$ of lane $l \in \mathcal{L}$ at time $t$, where $\mathcal{L}$ denotes the set of all approaching lanes to the intersection. Additionally, $\mathcal{L}^{in}$ and $\mathcal{L}^{out}$ refer to entering and exiting lanes, respectively, and so $\mathcal{L} = \mathcal{L}^{in} \bigcup \mathcal{L}^{out}$. Further, define *traffic movement* $v_l$ as a set that maps traffic of lane $l \in \mathcal{L}^{in}$ to possible leaving lanes $l' \in \mathcal{L}^{out}$. In this context, the set of valid traffic movement $p \in \mathcal{P}$ during a green light are called a *phase*. Define participating lanes $\mathcal{L}_p$ as the set of lanes that have appeared in at least one traffic movement of phase $p$. Note that each phase has a minimum time, and following that time, a decision about the next phase should be made.

The definition of the RL problem requires translating the domain into states, actions, and rewards. They are modeled utilizing the traffic characteristic $s_l^t$ as the state at time $t$. For actions, this is represented by the assigned active phase at time $t + 1$. Additionally, two attention mechanisms to define the policy $\pi^t$ to select the next action at time $t$ (i.e., the next phase). Specifically, the first attention mechanism are used defines weights to be used in considering the states of the corresponding lanes. This attention step, modeled in Equation (27), generates a vector of weights $w_p^t := \{w_l^t, l \in \mathcal{L}_p\}$, where each weight corresponds to a specific lane relevant for phase $p$ (i.e., $l \in \mathcal{L}_p$). The weights depend on a function ($r_p^t$) of the current states ($s^t$) of these lanes, and the average of the function across all relevant lanes, $q_p^t$. Intuitively, this attention allows focusing the mechanism on lanes depending on their closeness to the "average" (relevant) lane.

$$w_p^t = \texttt{state-attention}(r_p^t, q_p^t), \forall p \in \mathcal{P} \tag{27}$$

The weights are used to compute a representation of each phase $p$, as $z_p^t = \sum_{l \in \mathcal{L}_p} w_l^t \times g(s_l^t)$. The representations of the possible phases $p \in \mathcal{P}$ are then fed into an LSTM to capture the sequential dependence between phases. The output of the LSTM ($o^t$) is used, together with the representation of the possible phases $z_p^t$, $p \in \mathcal{P}$ for the second attention mechanism, modeled as in Equation (28). Specifically, this provides the probability $\pi^t := \{\pi_p^t, p \in \mathcal{P}\}$ of switching to any of the phases of the next timestep.

$$\pi^t = \texttt{action-attention}(\{z_p^t, p \in \mathcal{P}\}, o^t) \tag{28}$$

The reward is based on the negative of the intersection pressure defined in the work of Wei et al. [158]. The intersection pressure relates the lane capacity and the lane flow:

$$w(l, m) = \frac{v_l}{v_l^{max}} - \frac{v_m}{v_m^{max}} \tag{29}$$

for incoming and outgoing lanes $l$ and $m$, where $v_l^{max}$ refers to the maximum lane capacity for lane $l$. Note that this is an indication of the incoming and outgoing flow of traffic. Based on this paradigm, the RL agent learns a policy $\pi$ that suggests a phase for the next timestep for the current state $s^t$. This policy is learned based on the algorithm illustrated earlier and uses the cumulative rewards for policy updates. This provides the means to map states to actions in the AttendLight algorithm.

The AttendLight algorithm was tested using three-way and four-way intersections with varying numbers of lanes, phases, and flow rates. The states $s_l^t$ represent chunks $c \in \{1, 2, 3\}$ of the road leading up to the intersection for lane $l$ and are 100-m segments of a 300-m length of the lane. Each lane $l$ has a corresponding number of vehicles $\alpha_{l,c}^t$ in a chunk $c$ at time $t$. Further, lanes also may contain waiting vehicles and the quantity of waiting is represented by $\beta_l^t$. Therefore, the traffic characteristic for a lane is defined as $s_l^t := [\alpha_{l,1}^t, \alpha_{l,2}^t, \alpha_{l,3}^t, \beta_l^t]$. The proposed algorithm is tested against a number of reference baselines in the literature, showing a significant improvement in terms of lower Average Travel Time (ATT).

*5.3.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [42].

### 5.4 Game Theory

*5.4.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [97, 127]. Note that since Game Theory is a widely investigated and well-known topic, in the following we only sketch very briefly the key ideas behind this theory. The main reason for mentioning it in this survey is to place it in the overall context of human behavioral models. Additionally, we do not mention explicitly applications and results, as the literature is extensive and pinpointing only a (few) specific example(s) would be not that useful for the readers.

*5.4.2 Principles and Definitions.* Extending and utilizing the notion of rational behavior, Game Theory focuses on the interdependence of choices when the circumstances involve a collection of individuals [97]. In Game Theory, the decision maker, often referred to as the player, is operating with an assumed feedback signal. The feedback (payoff) provides a value associated with how desirable or costly an outcome might be for the player. Based on the payoff and anticipated behavior of other players, a player can attempt to optimize their choice (action) to best ensure an acceptable outcome. In this case, rather than attempting to learn a policy of behavior, the models of utility define the constraints of an optimization problem. Based on these constraints, the players of a game can find a suitable solution that follows the assumptions of rationality.

More formally, a player $i$ will have a set of available actions $A$ from which they can select an action $a_i$. The optimality of an action is dependent on multiple factors. First, the player has a payoff function $u : A \rightarrow \mathbb{R}$ that will map an action to a value. This mapping depends on the definition of the problem and the interdependence of actions to values between the players. The values $u(a)$ given by $u$ allow the player to generate a preferential ordering of actions, which indicates a player's need to identify desirable actions. Therefore, if the preferences of the player satisfy the following, then that player is considered rational under certainty [12]:

(1) (Completeness) $a_1 \succeq a_2$ or $a_2 \succeq a_1$;
(2) (Transitivity) If $a_1 \succeq a_2$ and $a_2 \succeq a_3$, then $a_1 \succeq a_3$.

The payoff values and the assumptions relating to rationality can then be utilized to identify an appropriate action or behavioral policy. In Game Theory, the behavioral policy of a player is referred to as the player's *strategy* or *strategy profile*.

The strategy of a player is a distribution over actions indicating the likelihood of selecting an action. A strategy that places the mass on more than one action would be considered a mixed strategy, which indicates that a player does not select a particular action 100% of the time in a given scenario (i.e., a pure strategy).

The identification of a strategy is the process of finding an equilibrium. In a Nash equilibrium, the player strategies are such that deviation from the current strategy for any player would be undesirable, as it would lead to another player having the means to take advantage of the change to achieve a better result. Depending on the game, it is possible to find zero, one, or even multiple equilibria. In the case of multiple equilibria, it is possible for the payoff for a particular player to vary between the equilibria, but it remains true that a deviation from the equilibrium from a single player would be undesirable [97].

Based on the problem definition, the approach for finding an equilibrium often comes down to an optimization problem. Given a set of $N$ players $\mathcal{N} = \{1, \ldots, N\}$ where each player $\nu$ has their own strategy $x^\nu \in \mathcal{X}_\nu \subseteq \mathbb{R}^{n_\nu}$, each player has an objective function $\theta_\nu(x^\nu, x^{-\nu})$ and constraints $g_i^\nu(x^\nu, x^{-\nu}) \leq 0, (i = 1, \ldots, m_\nu)$, which depend on their strategy and the strategy of others

$x^{-\nu} := (x^{\nu'})_{\nu' \neq \nu}$, an equilibrium in a Nash Game can be found as the solution to the optimization problem [76]:

$$P_\nu(x^{-\nu}) \quad \begin{array}{ll} \underset{x^\nu}{\text{minimize}} & \theta_\nu(x^\nu, x^{-\nu}) \\ \text{subject to} & g_i^\nu(x^\nu, x^{-\nu}) \leq 0, i = 1, \ldots, m_\nu \\ & x^\nu \in \mathcal{X}_\nu. \end{array}$$

The payoffs and assumption of rationality provide the constraints for an optimization-based solution method. However, the interdependency of the behaviors, outcomes, and payoffs can increase the difficulty of finding a suitable solution.

## 6 UNCERTAINTY AND IRRATIONALITY

In the following sections, we discuss topics relating to the modeling of uncertainty and biases. In Section 6.1, we discuss the use of quantum representations to accommodate for uncertainty and support a quantum representation of states. For Section 6.2, we discuss topics focusing on the resulting biases that come from the use of heuristics and similar shortcuts in reasoning. Further, these limitations in reasoning generate immediate and long-term effects, motivating studies on fairness.

### 6.1 Quantum Representations of Decisions and Irrational Thinking

*6.1.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [4, 32, 67, 90, 95].

*6.1.2 Principles and Definitions.* It has been argued that traditional probabilistic representations do not fully represent the reasoning of humans [94] or can require exponentially more complex representations [90, 95]. Instead, researchers have suggested the use of quantum-based methods for representing the statistical/probabilistic relationships between knowledge [67]. The argument is that the superposition-like representation better demonstrates how humans can have varying beliefs, which might not directly match the assumptions or requirements of probability (e.g., summing to 1). This allows for a representation in cases where reasoning operates in a state representing multiple possible outcomes or outcomes representing the same indefinite state. This is also a proposed method to account for potentially irrational or probability-violating reasoning of humans [32, 63]. The use of quantum representations also allows for replicating or modeling fallacies in human reasoning [4]. A common method for this representation is the use of a quantum-based Bayesian Network. In this case, a similar representation of the Bayesian Network is utilized, but the dynamics are represented using quantum representations of probabilities.

*Quantum Dynamics for Decision Models.* As defined in Section 3.1, sequential decision making can be modeled using an MDP. This representation provides a mechanism for representing the transition between world states based on a decision or action. Such a representation enables learning associations between actions and outcomes to build a model of behavior. Busemeyer et al. [21] a quantum dynamics model is proposed to represent brain processes as a replacement for the classic MDP model. The quantum representation enables transitioning from a "single path" assumption to one that represents unknown previous states as a "superposition of states." The superposition model allows for modeling interference effects for unobserved paths, which violates the Markov representation.

As noted by Busemeyer et al. [21], a key distinction in the representations is as follows:

> According to the Markov model, for any given realization, the unobserved system occupies exactly one basis state $|j\rangle$ at each moment in time. A sample path of the

> Markov process is a series of jumps from one basis state to another, which moves like a bouncing particle across time. A different path is randomly sampled for each realization of the Markov process. According to the quantum model, for any given realization, the unobserved system does not occupy any particular basis state at each moment in time. A realization of the quantum process is a fuzzy spread of membership across the basis states, which moves like a traveling wave across time. Each realization of the quantum process is identical, producing the same series of states across time. All of the randomness in the quantum model results from taking a measurement.

Hence, the quantum representation supports the notion of wave interference in the dynamics representation.

From a modeling standpoint, the main difference between a Markov and a Quantum representation is therefore the following. In a *specific realization* of a Markov process, the state at any point in time is deterministic and can be represented with $|P(t)\rangle$ as follows:

$$|P(t)\rangle = \sum_{j \in \Omega} w_j(t) \cdot |j\rangle, \tag{30}$$

where $j$ denotes the possible states, $\Omega$ is the set of possible states, and $w_j(t) \in \{0, 1\}$ is an indicator variable. Essentially, $|P(t)\rangle$ provides the (unique) state of the process at time $t$ for the specific realization.

On the contrary, in a quantum representation, the specific realization, at any point in time $t$ is not deterministically in one and only one state, but is in a *superposition* of all possible states, each with a given *weight*. In other words, even for a specific realization the process can be in *any* of the possible states, whereas the uncertainty can be only removed by "measuring" the specific state of the process. Specifically, the state, at time $t$ of the realization, denoted as $|\psi(t)\rangle$ is modeled as follows:

$$|\psi(t)\rangle = \sum \psi_j(t) \cdot |j\rangle, \tag{31}$$

where $\psi_j(t)$ is a complex number representing the probability amplitude that the process is in the specific state $|j\rangle$ at time $t$. As in any quantum superposition case, note that the squared magnitude of $\psi$ must be unity (i.e., $|\psi|^2 = 1$) to ensure the squared amplitudes produces a probability distribution over the possible states.

Finally, transitions between states become specific quantum operations over the representation of the states provided by $|\psi(t)\rangle$, which translate the traditional concept of transition probabilities between states of deterministic Markov processes.

*6.1.3 Applications and Recent Results.* He and Jiang [55], quantum representations are investigated in the context of categorization. The use of a quantum model is intended to represent interference effects observed in categorization and the resulting impact of decision making. The inclusion of a quantum system allows the modeling of a state that represents uncertainty in the reasoning process. This models how humans demonstrate hesitance when facing a decision. Two paradigm conditions were considered in the experiments: categorization decision making (C-D) and decision alone (D alone). In the C-D condition, participants were shown pictures of faces varying along two dimensions: face width and lip thickness. Participants were asked to categorize the face as a "good" ($G$) guy or a "bad" ($B$) guy and then make a decision to "attack" ($A$) or to "withdraw" ($W$). In the D alone condition, the participants were asked to make a decision directly without categorizing, but the faces shown were the same as in the C-D condition.

*Proposed Method.* Although categorization happens in the belief representation, it can influence the action part by producing the interference effect, which can also model the disjunction fallacy

(i.e., the false judgement that the probability $P(A \cup B)$ is less than either $P(A)$ or $P(B)$). Consequently, the proposed approach utilized a method to predict it. In the given problem space, the initial state involves six combinations of beliefs and actions:

$$\{|B_G A_A\rangle, |B_G A_U\rangle, |B_G A_W\rangle, |B_B A_A\rangle, |B_B A_U\rangle, |B_B A_W\rangle\}, \tag{32}$$

where, for example, $|B_G A_A\rangle$ symbolizes a participant categorizing the face as good while still intending to attack. Since participants are assumed to have some potential to be in any of the six quantum states, the person's state is a superposition of the six basis states

$$\begin{aligned} |\psi\rangle = &\psi_{AG} \cdot |B_G A_A\rangle + \psi_{UG} \cdot |B_G A_U\rangle + \psi_{WG} \cdot |B_G A_W\rangle \\ &+ \psi_{AB} \cdot |B_B A_A\rangle + \psi_{UB} \cdot |B_B A_U\rangle + \psi_{WB} \cdot |B_B A_W\rangle \end{aligned} \tag{33}$$

with initial state corresponding to an amplitude distribution

$$\psi(0) = \begin{bmatrix} \psi_{AG} \\ \psi_{UG} \\ \psi_{WG} \\ \psi_{AB} \\ \psi_{UB} \\ \psi_{WB} \end{bmatrix}, \tag{34}$$

where $|\psi_{XY}|^2$ is the probability of observing state $|B_X A_Y\rangle$ initially. As an assumption, the initial state is treated as equally distributed. In the process of decision making, updated information regarding a player's beliefs causes a transition in belief states. The decision maker must convert this transition in reasoning states into a decision. To convert the observations and measurements into actions, participants must convert the uncertain state $U$ to either $A$ or $W$. This represents when the decision maker transitions from uncertainty due to hesitation to a decision state. In this case, Pignistic Probability Transformation (PPT) was utilized. This provides the following total probability of attacking given that the face is categorized as $G$ or $B$, respectively:

$$P(A|G) = \left\| \Psi(A|G) + \frac{1}{2}\Psi(U|G) \right\|^2, \tag{35}$$

$$P(A|B) = \left\| \Psi(A|B) + \frac{1}{2}\Psi(U|B) \right\|^2, \tag{36}$$

where $\Psi(A|X)$ is the conditional amplitude (i.e., quantum equivalent of conditional operator) of attacking given the face is categorized as $X \in \{G, B\}$. The preceding denotes an even attribution of the probability of transitioning from the uncertain state to decision. This then provides the means to calculate the total probability of attacking:

$$\begin{aligned} P(A) &= P(G)P(A|G) + P(B)P(A|B) \\ &= \sum_{X \in \{G,B\}} \left( |\psi_{AX}|^2 + |\psi_{UX}|^2 + |\psi_{WX}|^2 \right) \cdot \left\| \Psi(A|X) + \frac{1}{2}\Psi(U|X) \right\|^2. \end{aligned} \tag{37}$$

A similar approach is utilized for $P(W)$ and in the D alone condition. This allows a model of representing the state of uncertainty encountered before a decision is made and the process of converting updated beliefs into an action. Based on this model, the action likelihoods can be represented and estimated using a quantum modeling approach.

For experiment conditions, the proposed method was tested against prior observations of human-generated data as well as prior prediction model results. In the data generation process, 26 participants were asked to categorize the face as *good guy* ($G$) or *bad guy* ($B$) and then make a decision to *attack* ($A$) or to *withdraw* ($W$). The faces roughly fall into two categories: "narrow"

faces (narrow width and thick lips) or "wide" faces (wide width and thin lips). Participates were informed that "narrow" faces had a 0.6 probability of belonging to the "bad guy" population. Similarly, participants were informed that "wide" faces had a 0.60 probability of belonging to the "good guy" population. Rewards were given for choosing attack for *B* and withdraw for *G*. In the D alone condition, the participants were asked to make a decision directly without categorizing, but the faces shown to participants were the same as those in the C-D condition. Each participant provided 51 observations for the C-D condition and 17 observations for the D alone condition. The proposed approach showed a strong alignment with the pattern of behavior observed in the human participant results.

Performance of the proposed approach with respect to sensitivity analysis was also demonstrated. The experiment outcomes show how sensitive the results of the method are to a ±5% change in the original human results. In this test, the method again shows strong performance, but results indicated a stronger sensitivity to changes in the *B* case versus the *G* case. For a plot of these results, please refer to the original article [55].

*6.1.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [42].

### 6.2 Biases and Fairness in Representations and Understanding

*6.2.1 Relevant Survey(s).* For relevant survey papers and related topics, please refer to other works [19, 29, 29, 45, 73, 84, 89, 104, 124, 141].

*6.2.2 Principles and Definitions.*

*Bias.* Humans demonstrate multiple forms and sources of bias. In one case, humans can demonstrate skewed interpretations of data, probabilities, confidence, and so forth. [84]. This kind of irrationality and bias is often attributed to the use of cognitive heuristics. It is argued that the use of heuristics causes humans to often accept coarse analysis and suboptimal solutions. This demonstrates how humans can violate the traditional notions of rational behavior. Further, the use of heuristics and other shortcuts can lead to skewed interpretations of information. These biases and coarse representations can lead to sensitivity regarding the interpretation of rare events, risk, probability, and more. Related, it has also been shown how humans tend to under-react to probabilistic information and also fail to follow belief updates modeled by Bayes' rule [29].

Another aspect of bias can come in the form of inattentiveness, which could be linked to the bounded computational power of human cognition. Humans also demonstrate a resistance to changes in views or opinions when facing contradictory information [89]. In fact, it is common for people to become more deeply convicted in their views rather than convinced they might have been wrong. Bias can also come in the form of inductive bias introduced in architecture design and algorithmic choices made by the humans generating algorithms. Both implicitly and explicitly, humans introduce inductive biases into the artificial systems they are developing. In fact, Goyal and Bengio [47] note how much of the success of Deep Learning models could be attributed to the inclusion of inductive bias in these systems. Further, how the use of inductive biases might be a requirement for the creation of generalized artificial intelligence. In [47], it is noted how biases allow for assumptions regarding the problem being solved or the interpretation of information, which makes the system better suited for adaptation to broader datasets.

*Probability-Based Models of Behavior and Reasoning.* To model deficiencies or biases in understanding and reasoning in humans, researchers have investigated Bayesian representations and related topics. For instance, (Hierarchical) Bayesian models can demonstrate how humans reason about likelihoods of outcomes/scenarios [29, 30, 119, 144, 148, 172], and how human reasoning

shows flawed interpretations or skewed scales of importance (high or low). This illustrates how humans can under-react to probabilities [29, 30] or estimate likelihoods based on observed frequencies. Studies have shown how people tend to select options at a similar frequency to their probability of good outcome rather than developing a bias to the choice with the highest likelihood of a good outcome [172]. Similarly, Bayesian models can replicate the performance of human study participants when tasked with selecting in a multi-arm bandit problem [122]. This type of representation of understanding can also be extended to a causal model, which models the cause and effect relationship between different environment aspects [51]. Causal models are an important aspect of human reasoning as they allow for predictions and retrospective reasoning.

*Fairness.* With the use of algorithms, a consideration needs to be made with respect to the fairness of outcomes. As noted in the work of Lee [82], fairness requires equal or equitable treatment of everyone based on their performance or needs. As a concept, this is straightforward. There should be as little (preferably none) bias as possible in the treatment of individuals. We would prefer the outcome of a human's decision be fair, so it is natural to desire the same behavior from an algorithm. Fairness can be a quantity measured with respect to a metric [57], or fairness can represent a qualitative impression people have about a feature or outcome [49, 82]. These measures provide an indication of how people perceive the features of an algorithm or its behavior. The notion of fairness can also be considered with respect to longer timescales and deeper consequences. In the work of Heidari et al. [57], the longer effects of fairness measures and decisions are weighed to indicate how strongly an approach may impact the behavior of the humans affected.

*6.2.3 Applications and Recent Results.* Heidari et al. [57], population-level changes at the macro scale that are caused by algorithmic decisions and how this relates to fairness are modeled. Using measures of segregation from sociology and economics, the proposed models quantified these changes. This allowed measuring different directions of shift in the group-conditional distribution based on the different models demonstrated. In this context, [57] notes how most notions or measures of fairness assume a static population. It is argued that this approach fails to account for long-term welfare and prosperity. Therefore, the proposed measure considers a notion of *effort* based on economic literature on Equality of Opportunity. The effort function highlights the idea that the necessary changes for a desirable outcome often are more difficult for a member of a disadvantaged group compared to an advantage counterpart. Based on this concept, *effort unfairness* is formulated as the discrepancy in effort required for members in different groups to obtain desired outcomes.

In the proposed approach, individuals are assumed to imitate an exemplar individual who demonstrates a more desirable algorithm outcome, which is related to social learning—the assumption being that the observer would believe this imitation offers a higher likelihood of a better outcome, which suggests an individual would exert effort to attain a replication of the exemplar's social model if doing so could result in higher overall utility. To model this dynamic, the approach uses a group-dependent, data-driven measure of effort, which is inspired by the literature on Equality of Opportunity. This is noted as the effort it takes individual $i$ to improve their feature $k$ value from $x$ to $x'$. This effort is assumed proportional to the difference between the rank/quantile of $x'$ and $x$ in the distribution of feature $k$ in $i$'s social group. An individual who successfully replicates the role model obtains a positive utility (reward minus effort).

In this context, the approach utilizes a standard supervised learning setting with training data set $D = \{(x_i, y_i)\}_{i=1}^{n}$ of $n$ instances $z_i = (x_i, y_i)$, where $x_i \in \mathcal{X}$ specifies the feature vector for individual $i$ and $y_i \in \mathcal{Y}$, the ground truth label for him/her. Further, let $s_i$ refer to the sensitive feature value (e.g., race, gender, or their intersection) for individual $i$. To measure the impact of a

change in label, define benefit function $b : \mathcal{X}, \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ that quantifies the benefit of a change in label from $y$ to $\hat{y}$ and is assumed a linear function.

To formally define measures of effort, reward, and utility, refer to the following definitions. For deployed model $h$ and individual characterized by $z = (x, y)$, let $\mathcal{R}_h(z, z')$ specify the reward/benefit as a result of characteristic change from $z$ to $z'$:

$$\mathcal{R}_h(z, z') = b(h(x'), y')^{\alpha} - b(h(x), y)^{\alpha}, \tag{38}$$

where $\alpha > 0$ is a constant specifying the individual's degree of risk aversion. Again, this demonstrates a measure that considers the impact on an individual's outcome when they change a characteristic. For instance, a person could attain a degree or find a new job. Let $\mathcal{E}_h(z, z')$ specify the effort required for a qualification change from $z$ to $z'$. The overall utility of the individual is then $\mathcal{U}_h(z, z') = \mathcal{R}_h(z, z') - \mathcal{E}_h(z, z')$ (i.e., utility equals reward minus effort).

Based on the preceding assumptions and definitions, [57] defines several effort-based measures of unfairness, including the following.

*Definition 6.1 (Effort-Reward Unfairness).* For a predictive model $h$, the effort-reward unfairness is the inequality of the following metric across different groups:

$$\mathbb{E}_{q \sim P : s_i = s} \max_{z \in \mathcal{Z}} \mathcal{U}_h(z_i, z). \tag{39}$$

This provides a measure of the performance of the highest-utility members of a group achieved by a higher level of effort. The impact of both threshold $\delta$ and the models are noted. It is demonstrated that $\delta$ and the model can cause a significant change in outcome and effort level for the same data.

To measure performance, several techniques from sociology measuring segregation were utilized. First, *evenness* measures how unevenly the minority group is distributed over areal units. The evenness value is maximized when all units have equivalent relative numbers of minority and majority members as the whole population. For a formal definition, the proposed approach used the Atkinson Index, which measures inequality. The second measure, *centralization*, is the degree to which a group is spatially located near the center of an urban area. This can be measured by comparing the percentage of minority residents living in the central areas of the city. Further, the approach utilized the Centralization Index, which is defined as

$$\frac{\sum_{i \text{ central}} m_i}{m}, \tag{40}$$

where $m$ is the total minority population. The third measure of segregation, the Absolute Clustering Index, "expresses the average number of [minority] members in nearby [areal units] as a proportion of the total population in those nearby [areal units]."

To demonstrate the impact of differing measures and concepts of fairness, the primary cases of interest were that of linear regression. The model was trained by minimizing the mean squared error while imposing welfare constraints defined in the work of Heidari et al. [56]. The results demonstrate the impact of the fairness constraints on the three segregation measures defined earlier when females are the minority group considered. It is noted that the expectation would be such that the constraints would result in reduced segregation in the long run, but the results demonstrate deviations. Let $\tau$ denotes a minimum threshold on $\mathcal{U}$ (i.e., $\mathcal{U} \geq \tau$) with respect to learning with the loss function $L_D$. The results show that in the case of small $\tau$ values, enforcing fairness constraints can generate a reduction in the degree of clustering. Conversely, larger values of $\tau$ can generate the opposite effect. Regarding evenness, this measure is relatively unchanged for all tested values of $\tau$. As noted in [56], the results indicate (1) a label of *desirable* for some members of the disadvantaged group motivates those members to remain unchanged, and (2) the positively

labeled members can serve as role models for other group members and motivate positive changes in others.

*6.2.4 Additional Relevant Results.* For further examples of relevant results, please refer to the work of Fuchs et al. [42].

## 7 CONCLUSION

In this article, we demonstrated methods focusing on multiple aspects of human behavior and cognition in addition to how humans interact with artificial systems. The types and end uses for these interactions vary, but a key concept is the ability to learn from or adapt to humans. The pervasive diffusion of AI systems provides exceptional opportunities to build next-generation autonomous and adaptive systems in several application areas (which we have described briefly in Section 2). To gain full advantage of this opportunity, it is of paramount importance that AI systems "learn to know and anticipate" the behavior of the involved users. This is a cornerstone of HCAI systems for autonomous and adaptive behaviors, which require embedding practical models of the human behavior that can be used to interpret users' actions and anticipate their reactions with respect to (certain or likely) stimuli. Humans use multiple advanced skills daily to navigate the world and their interactions with others. Researchers have investigated several topics such as ToM, IRL, AL, and more in an attempt to capture some of these capabilities and imbue systems with more advanced capabilities. This enables systems to better understand and operate in the world, including cases in which there are humans to account for and support, or when systems are expected to operate autonomously. Further, it enables systems to perform tasks more skillfully by replicating the level of skill demonstrated by humans. In addition to learning from humans, these techniques often serve to enable richer and more intuitive interactions between humans and artificial systems. This enables both the interactions as well as the system's ability to account for the human and potentially gain further knowledge as a result of the interaction. This approach gives systems a way in which to account for the cyclical nature of these relationships to improve performance and learn.

### 7.1 Relevant Pros and Cons

In this section, we discuss some of the relevant benefits and detractors for the topics presented in this article. The following details will not be comprehensive, but instead provide preliminary impressions. These impressions are meant to guide readers in their effort to determine suitable methods for their own purposes.

*7.1.1 Learning Human Behaviors by Experience and Feedback.*

- *Reinforcement Learning*: Pros
  - Learns from exploration (i.e., trial and error) based on behavior-related feedback to learn behavior best fitting the definition of desired behavior.
  - Support for discrete and continuous representations of actions and environment.
- *Reinforcement Learning*: Cons
  - Exploration can be inefficient/expensive (i.e., can require large number of training episodes).
  - Learning sensitive to feedback definition (i.e., agent behavior can be unexpected if the reward is inadequately defined).
- *Inverse Reinforcement Learning and Imitation Learning*: Pros
  - Behavior learned via given examples of desirable behavior.
  - For IRL, an estimation of the underlying feedback motivating behavior is generated.

  - For IL, the estimation of the feedback is omitted, reducing the complexity of the learning method.
- *Inverse Reinforcement Learning and Imitation Learning*: Cons
  - Requires access to examples of behavior and a sufficient-sized dataset to help avoid overfitting.
  - In IRL, a two-step process (estimating feedback and policy learning) increases the complexity of the learning process.
- *Active Learning*: Pros
  - Access to a teacher for representation of desirable behavior/labels for the learning method.
  - Reduction in size of labeled data as the learner is motivated to prioritize input with lower confidence regarding accuracy of the model (i.e., higher "curiosity" for items of lowest confidence).
  - Reduced exploration complexity via teacher feedback/guidance.
- *Active Learning*: Con
  - Requires access to a reliable teacher and accurate labels.

#### 7.1.2 *Belief and Reasoning Approaches.*

- *Meta-reasoning and Meta-learning*: Pros
  - Agents learn to map context/task to a suitable model/representation to guide behavior.
  - Enables constraints on resource usage to *learn when to learn* and when to use shortcuts (e.g., heuristics).
- *Meta-reasoning and Meta-learning*: Cons
  - High complexity of learning (need to learn a mapping between the context and behavior model as well as learning the behavior model).
  - Forgetting can occur as a result of new experiences overcoming previously learned knowledge.
- *Theory of Mind*: Pros
  - Enables estimation of private information and mental states of others (including recursive levels of representation to model the view of self from another perspective).
  - Performance improvement given an accurate model of ToM via estimated anticipated behavior of others.
- *Theory of Mind*: Cons
  - Can require an extensive amount of training.
  - Can rely on increased levels of domain knowledge to model the mental state of others.
- *Simulating Human Knowledge of World for Learners*: Pros
  - Replaces the need for an agent to learn the model of world dynamics to reduce learning complexity and improve planning.
  - Improves accuracy/realism of the world model in representation of an expected environment.
- *Simulating Human Knowledge of World for Learners*: Con
  - Models can be challenging and/or costly to define.

#### 7.1.3 *Bounded Rationality and Cognitive Limitations.*

- *Cognitive Heuristics*: Pros
  - Model inaccuracies and shortcuts defined based on observations of human reasoning.
  - Heuristics/Rules can be easier to implement and use than learned models.
- *Cognitive Heuristics*: Con
  - Rely on domain knowledge and potentially coarse representations of human reasoning.

- *Cognitively/Biologically Plausible Representations*: Pros
  - Provide a modeling system to portray aspects of human reasoning/behavior.
  - Can mimic cognitive limitations, biases, and so forth found in humans.
- *Cognitively/Biologically Plausible Representations*: Cons
  - Many of the popular models rely on extensive/expensive world models.
  - Can require large numbers of training samples in the learning case.
- *Attention*: Pros
  - Representation of how humans ignore or reduce the significance of stimuli that can support Deep Learning methods.
  - Models can learn relationships across various input values to learn deeper associations.
- *Attention*: Con
  - As with many Deep Learning scenarios, these methods can be difficult to train, require a large amount of samples, result in highly complex networks, and so forth.
- *Game Theory*: Pros
  - Incorporates assumed rational behavior to model outcomes.
  - Identify appropriate behavior based on the payoff received corresponding to selected actions.
- *Game Theory*: Cons
  - Relies on conservative assumptions regarding human behavior (not all humans behave rationally).
  - Representation of more realistic scenarios can have high levels of complexity.

#### 7.1.4 *Uncertainty and Irrationality.*

- *Quantum Representations of Decisions and Irrational Thinking*: Pros
  - Enables rich models of uncertainty including those that would violate traditional probability rules (e.g., summing to 1).
  - Model fallacies in reasoning (e.g., disjunction fallacy).
  - Supports modeling of an interference effect for unobserved paths in Markov models rather than assuming a single path.
- *Quantum Representations of Decisions and Irrational Thinking*: Cons
  - Model complexity when compared to traditional models.
  - Access to quantum computing resources or systems that can represent the quantum models.
- *Biases and Fairness in Representations and Understanding*: Pros
  - Incorporates models representing human biases in reasoning, interpretation of information, and so forth.
  - Enables considerations regarding effort, unfairness, biases, and so forth for decision models.
- *Biases and Fairness in Representations and Understanding*: Con
  - Costly and difficult to generate models.

### 7.2 Complementary Concepts

In this section, Table 1 denotes our interpretation of the complementarity of the various topics described in this article. As this is primarily a qualitative analysis, there will of course be a level of subjectivity to the assessment. Therefore, the table should serve as a preliminary indication of relation, compatibility, and so forth for the various topics. We will not discuss all relationships presented in the table, but instead will provide particular examples to illustrate our reasoning. For instance, the significant number of entries relating RL to other topics can be attributed to

Table 1. Levels of Complementarity: $<$ (Low), $\ll$ (Medium), $\lll$ (High)

| Topic | 3.1 | 3.2 | 3.3 | 4.1 | 4.2 | 4.3 | 5.1 | 5.2 | 5.3 | 5.4 | 6.1 | 6.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.1: Reinforcement Learning | – | $\lll$ | $\lll$ | $\ll$ | $\ll$ | $\lll$ | $<$ | $\ll$ | $\ll$ | $<$ | $<$ | |
| 3.2: Inverse Reinforcement Learning and Imitation Learning | $\lll$ | – | $\ll$ | $<$ | | | | | | | | $\ll$ |
| 3.3: Active Learning | $\lll$ | $\ll$ | – | $<$ | $<$ | | | | | | | $\ll$ |
| 4.1: Meta-reasoning and Meta-learning | $\ll$ | $<$ | $<$ | – | | | $\ll$ | $\ll$ | | | | |
| 4.2: Theory of Mind | $\ll$ | | $<$ | | – | | | $<$ | | | | |
| 4.3: Simulating Human Knowledge of World for Learners | $\lll$ | | | | | – | | $\ll$ | | | | |
| 5.1: Cognitive Heuristics | $<$ | | | $\ll$ | | | – | $\ll$ | $\ll$ | | | $\ll$ |
| 5.2: Cognitively/Biologically Plausible Representations | $\ll$ | | | $\ll$ | $<$ | $\ll$ | $\ll$ | – | | | | |
| 5.3: Attention | $\ll$ | | | | | | $\ll$ | | – | | | |
| 5.4: Game Theory | $<$ | | | | | | | | | – | | |
| 6.1: Quantum Representations of Decisions and Irrational Thinking | $<$ | | | | | | | | | | – | |
| 6.2: Biases and Fairness in Representations and Understanding | | $\ll$ | $\ll$ | | | | $\ll$ | | | | | – |

the fact that several of these topics rely on similar notions of relating behavior/determinations to a feedback signal denoting desirable outcomes/behavior. Further, utilizing concepts from ToM, Attention, and so forth enables the use of RL in more complex environments, with more human-like ingestion of input, compensation for others, and more. In the case of AL, we can see that there is indeed a motivation to consider topics such as biases in reasoning. For example, the bias of behavior exhibited by the teacher can have a significant impact on the outcomes with respect to the learning process. With respect to cognitive architectures and related biologically plausible systems, it is apparent that the inclusion of simulations can prove useful. The assistance of simulations allows systems to more accurately mode the dynamics of the world to replicate how a human might predict physical motion. This enables richer modeling from the cognitive architecture and allows for more true-to-life modeling/reasoning.

### 7.3 Relevant Open Problems and Challenges

In the following sections, we discuss general aspects regarding several topics presented in this article and how shortcomings of current approaches motivate future work. This is not meant as a comprehensive list, but rather examples of topics for future consideration.

#### *7.3.1 Reinforcement Learning.*

*Dynamic Programming Algorithms.* In RL, many methods rely on learning methods and algorithms utilizing Dynamic Programming. According to Levine et al. [83] and Prudencio et al. [112], poor performance can occur in the offline settings due to issues such as distributional shift—the underlying cause being the distributional shift of the actions due to the discrepancy between the behavior policy and the current learned policy. It is noted that policy constraints and explicit uncertainty estimation can help mitigate these issues, but these are not without their own shortcomings. Further, as is seen in Deep Learning and related topics, limited access to training data can result in overfitting, brittle representations, and a lack of generalizable behavior. Similarly, methods can further struggle when they encounter out-of-distribution states. This can impact various aspects of the underlying model, including the state densities. Consequently, the resulting output of the

Dynamic Programming algorithms can have high inaccuracy in the case of these lower probability states. According to Prudencio et al. [112], a remaining issue is the lack of realistic scenarios and access to corresponding data. This can significantly reduce the accuracy of the representation and its suitability for the real world. Related, the generation of large amounts of accurately labeled is still quite challenging and costly. As such, the access to realistic and accurate data remains a concern.

*7.3.2 Inverse Reinforcement Learning and Imitation Learning.*

*Inverse Reinforcement Learning.* According to Arora and Doshi [9], IRL suffers from ambiguity regarding valid solutions. This is due to the fact that multiple representations of reward can result in a representation explaining the sample data. Similarly, the accuracy of the resulting model can be measured in several ways. These measures of accuracy consider/prioritize different aspects: policy, value, and so forth. The method for defining accuracy will of course then determine which aspects are considered significant. For instance, measuring the variance in policies could result in a small divergence despite the difference occurring in a critical state. Therefore, careful consideration is needed when determining which details are measured regarding the accuracy. With respect to IL, Zheng et al. [170] provide multiple examples of open challenges: diverse behavior learning, suboptimal demonstrations, finding globally optimal solutions (as opposed to locally optimal), and so forth.

*7.3.3 Meta-learning and Meta-reasoning.* As indicated in the work of Huisman et al. [64], some relevant concerns regarding Meta-learning and Meta-reasoning include topics such as overfitting, access to realistic domains/data, and fair comparison of methodologies. In the case of overfitting, this follows many related topics as many examples of Meta-learning and Meta-reasoning are based on Deep Learning approaches, so they are susceptible to the same overfitting concerns as seen in Supervised Learning and other related topics. Regarding the comparison of methodologies, [64] indicates how different techniques can utilize different underlying backbones. Consequently, it is more challenging to distinguish whether the variance in performance is indeed methodological.

*7.3.4 Cognitive Architectures.* Regarding cognitive architectures, Kotseruba and Tsotsos [78] demonstrate several relevant open challenges. For instance, the importance of perception and the lack of systems that effectively support perception in observation and reasoning. Additionally, it is argued that many approaches instead rely on simulations to replace perception aspects, such as vision. Further, [78] notes how biologically inspired models commonly are unable to match the level of range and efficiency needed for practical applications as compared to those based on heuristics and related methods.

*7.3.5 Attention.* For topics related to Attention, Niu et al. [102] demonstrate several concepts of interest. Regarding the concepts of the query and key in self-attention, some work has demonstrated how these can be effectively combined. Therefore, further research should be performed to determine the importance of independent or combined representations for keys and queries. Additionally, there is still room for improvement regarding the computational complexity of attention models. As well, [102] notes the desirability of further investigations into the translation of techniques across use cases—for instance, applying a self-attention method from Natural Language Processing in a Computer Vision domain. As a final example, we can also note the need for effective performance indicators to better demonstrate the utility of various attention methods.

### 7.4 Concluding Remarks

Overall, we believe that the approaches and goals presented motivate the topics discussed in this article and serve to promote further investigation into how humans and artificial systems interact

and learn from each other. We can conclude there is not a one-size-fits-all approach to best model the human behavior in HCAI. Each approach has its own merits and deficiencies, and the best choice largely depends on the specific problem at hand. However, the literature on this topic is quite vast, which allows designers of HCAI systems to leverage an extensive toolbox to equip AI agents with practical approaches to model human behavior. Therefore, we believe that this area is going to emerge as one of the most active in the coming years, cutting across autonomous and adaptive systems, pervasive environments, and advanced AI agents.

## REFERENCES


[1] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*. 1.

[2] Rakefet Ackerman and Valerie A. Thompson. 2017. Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences* 21, 8 (2017), 607–617.

[3] Malik Ghallab, Adele Howe, Craig Knoblock, Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, et al. 1998. *PDDL—The Planning Domain Definition Language*. Technical Report CVC TR-98-003/DCS TR-1165. Yale Center for Computational Vision and Control.

[4] Diederik Aerts, Massimiliano Sassoli De Bianchi, Sandro Sozzo, and Tomas Veloz. 2021. Modeling human decision-making: An overview of the Brussels quantum approach. *Foundations of Science* 26, 1 (2021), 27–54.

[5] Mete Akbulut, Erhan Oztop, Muhammet Yunus Seker, X. Hh, Ahmet Tekden, and Emre Ugur. 2021. ACNMP: Skill transfer and task extrapolation through learning from demonstration and reinforcement learning via representation sharing. In *Proceedings of the Conference on Robot Learning*. 1896–1907.

[6] Kelsey R. Allen, Kevin A. Smith, and Joshua B. Tenenbaum. 2020. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences* 117, 47 (2020), 29302–29310.

[7] John R. Anderson. 2013. *The Adaptive Character of Thought*. Psychology Press.

[8] John R. Anderson, Lynne M. Reder, and Christian Lebiere. 1996. Working memory: Activation limitations on retrieval. *Cognitive Psychology* 30, 3 (1996), 221–256.

[9] Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297 (2021), 103500.

[10] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38.

[11] Alireza Asgari and Yvan Beauregard. 2021. Brain-inspired model for decision-making in the selection of beneficial information among signals received by an unpredictable information-development environment. Preprint (2021).

[12] Gholamreza Askari, Madjid Eshaghi Gordji, and Choonkil Park. 2019. The behavioral model and game theory. *Palgrave Communications* 5, 1 (2019), 1–8.

[13] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

[14] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[15] Thierry Baron, Martin D. Levine, and Yehezkel Yeshurun. 1994. Exploring with a foveated robot eye system. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Volume 2—Conference B: Computer Vision and Image Processing*. IEEE, Los Alamitos, CA, 377–380.

[16] Raunak Bhattacharyya, Blake Wulfe, Derek Phillips, Alex Kuefler, Jeremy Morton, Ransalu Senanayake, and Mykel Kochenderfer. 2020. Modeling human driving behavior through generative adversarial imitation learning. *arXiv preprint arXiv:2006.06412* (2020).

[17] Francesca Bianco and Dimitri Ognibene. 2019. Functional advantages of an adaptive theory of mind for robotics: A review of current architectures. In *Proceedings of the 2019 11th Computer Science and Electronic Engineering Conference (CEEC'19)*. IEEE, Los Alamitos, CA, 139–143.

[18] Boyan Paskalev Bontchev, Valentina Terzieva, and Elena Paunova-Hubenova. 2021. Personalization of serious games for learning. *Interactive Technology and Smart Education* 18, 1 (2021), 50–68.

[19] Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jon Lenchner, Nick Linck, Andrea Loreggia, et al. 2020. Thinking fast and slow in AI. *arXiv preprint arXiv:2010.06002* (2020).

[20] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *arXiv:1606.01540* (2016).

[21] Jerome R. Busemeyer, Zheng Wang, and James T. Townsend. 2006. Quantum dynamics of human decision-making. *Journal of Mathematical Psychology* 50, 3 (2006), 220–241.

[22] Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. StylePredict: Machine theory of mind for human driver behavior from trajectories. *arXiv preprint arXiv:2011.04816* (2020).
[23] Chiho Choi, Srikanth Malla, Abhishek Patil, and Joon Hee Choi. 2019. DROGON: A trajectory prediction model based on intention-conditioned behavior reasoning. *arXiv preprint arXiv:1908.00024* (2019).
[24] Andrzej Cichocki and Alexander P. Kuleshov. 2021. Future trends for human-AI collaboration: A comprehensive taxonomy of AI/AGI using multiple intelligences and learning styles. *Computational Intelligence and Neuroscience* 2021 (2021), 8893795.
[25] Marco Conti, Andrea Passarella, and Sajal K. Das. 2017. The Internet of People (IoP): A new wave in pervasive mobile computing. *Pervasive and Mobile Computing* 41 (2017), 1–27. https://doi.org/10.1016/j.pmcj.2017.07.009
[26] Alana de Santana Correia and Esther Luna Colombini. 2021. Attention, please! A survey of neural attention models in deep learning. *arXiv preprint arXiv:2103.16775* (2021).
[27] Stefania Costantini. 2002. Meta-reasoning: A survey. In *Computational Logic: Logic Programming and Beyond*. Springer, 253–288.
[28] Edward A. Cranford, Cleotilde Gonzalez, Palvi Aggarwal, Sarah Cooney, Milind Tambe, and Christian Lebiere. 2020. Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science* 12, 3 (2020), 992–1011.
[29] Ishita Dasgupta. 2020. *Algorithmic Approaches to Ecological Rationality in Humans and Machines*. Ph. D. Dissertation. Harvard University.
[30] Ishita Dasgupta, Eric Schulz, Joshua B. Tenenbaum, and Samuel J. Gershman. 2020. A theory of learning to infer. *Psychological Review* 127, 3 (2020), 412.
[31] Daniel de Almeida Rocha and Julio Cesar Duarte. 2019. Simulating human behaviour in games using machine learning. In *Proceedings of the 2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames'19)*. IEEE, Los Alamitos, CA, 163–172.
[32] Shahram Dehdashti, Lauren Fell, and Peter Bruza. 2020. On the irrationality of being in two minds. *Entropy* 22, 2 (2020), 174.
[33] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2021. The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354* (2021).
[34] Cvetomir Dimov, Patrick H. Khader, Julian N. Marewski, and Thorsten Pachur. 2020. How to model the neurocognitive dynamics of decision making: A methodological primer with ACT-R. *Behavior Research Methods* 52, 2 (2020), 857–880.
[35] Andrew D. M. Dobson, Emiel De Lange, Aidan Keane, Harriet Ibbett, and E. J. Milner-Gulland. 2019. Integrating models of human behaviour between the individual and population levels to inform conservation interventions. *Philosophical Transactions of the Royal Society B* 374, 1781 (2019), 20180053.
[36] M. Dolfin, L. Leonida, and N. Outada. 2017. Modeling human behavior in economics and social science. *Physics of Life Reviews* 22 (2017), 1–21.
[37] Malin Eiband, Daniel Buschek, and Heinrich Hussmann. 2021. How to support users in understanding intelligent systems? Structuring the discussion. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 120–132.
[38] P. A. Facione and C. A. Gittens. 2012. *Think Critically*. Pearson. https://books.google.it/books?id=YGM5ygAACAAJ.
[39] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2020. Deep inverse reinforcement learning for behavior prediction in autonomous driving: Accurate forecasts of vehicle motion. *IEEE Signal Processing Magazine* 38, 1 (2020), 87–96.
[40] Ismael T. Freire, Xerxes D. Arsiwalla, Jordi-Ysard Puigbò, and Paul Verschure. 2019. Modeling theory of mind in multi-agent games using adaptive feedback control. *arXiv preprint arXiv:1905.13225* (2019).
[41] Andrew Fuchs, Andrea Passarella, and Marco Conti. 2022. Modeling human behavior part I—Learning and belief approaches. *arXiv preprint arXiv:2205.06485* (2022).
[42] Andrew Fuchs, Andrea Passarella, and Marco Conti. 2022. Modeling human behavior part II—Cognitive approaches and uncertainty. *arXiv preprint arXiv:2205.06483* (2022).
[43] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614* (2021).
[44] Samuel J. Gershman, Eric J. Horvitz, and Joshua B. Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 6245 (2015), 273–278.
[45] Gerd Gigerenzer. 2008. Moral intuition = fast and frugal heuristics? In *Moral Psychology*. MIT Press, Cambridge, MA, 1–26.
[46] Cleotilde Gonzalez, Javier F. Lerch, and Christian Lebiere. 2003. Instance-based learning in dynamic decision making. *Cognitive Science* 27, 4 (2003), 591–635.

[47] Anirudh Goyal and Yoshua Bengio. 2020. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091* (2020).
[48] Michael S. A. Graziano. 2019. Attributing awareness to others: The attention schema theory and its relationship to behavioural prediction. *Journal of Consciousness Studies* 26, 3-4 (2019), 17–37.
[49] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. 903–912.
[50] Thomas L. Griffiths, Frederick Callaway, Michael B. Chang, Erin Grant, Paul M. Krueger, and Falk Lieder. 2019. Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences* 29 (2019), 24–30.
[51] Thomas L. Griffiths and Joshua B. Tenenbaum. 2005. Structure and strength in causal induction. *Cognitive Psychology* 51, 4 (2005), 334–384.
[52] Jürgen Groeneveld, Birgit Müller, Carsten M. Buchmann, Gunnar Dressler, Cheng Guo, Niklas Hase, Falk Hoffmann, et al. 2017. Theoretical foundations of human decision-making in agent-based land use models—A review. *Environmental Modelling & Software* 87 (2017), 39–48.
[53] Fatih Gurcan, Nergiz Ercil Cagiltay, and Kursat Cagiltay. 2021. Mapping human–computer interaction research themes and trends from its existence to today: A topic modeling-based review of past 60 years. *International Journal of Human–Computer Interaction* 37, 3 (2021), 267–280.
[54] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107.
[55] Zichang He and Wen Jiang. 2018. An evidential dynamical model to predict the interference effect of categorization on decision making results. *Knowledge-Based Systems* 150 (2018), 139–149.
[56] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems* 31 (2018), 1–12.
[57] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. 2019. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). Proceedings of Machine Learning Research, Vol. 97. PMLR, 2692–2701. https://proceedings.mlr.press/v97/heidari19a.html.
[58] Andreas Holzinger, Matthias Dehmer, Frank Emmert-Streib, Rita Cucchiara, Isabelle Augenstein, Javier Del Ser, Wojciech Samek, Igor Jurisica, and Natalia Díaz-Rodríguez. 2022. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion* 79 (2022), 263–278.
[59] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M. Pintea, and Vasile Palade. 2019. Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Applied Intelligence* 49, 7 (2019), 2401–2414.
[60] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2022. Explainable AI methods—A brief overview. In *Proceedings of the International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. 13–38.
[61] Andreas T. Holzinger and Heimo Muller. 2021. Toward human–AI interfaces to support explainability and causability in medical AI. *Computer* 54, 10 (2021), 78–86.
[62] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439* (2020).
[63] Zhiming Huang, Lin Yang, and Wen Jiang. 2019. Uncertainty measurement with belief entropy on the interference effect in the quantum-like Bayesian networks. *Applied Mathematics and Computation* 347 (2019), 417–428.
[64] Mike Huisman, Jan N. Van Rijn, and Aske Plaat. 2021. A survey of deep meta-learning. *Artificial Intelligence Review* 54, 6 (2021), 4483–4541.
[65] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys* 50, 2 (2017), 1–35.
[66] Joshua Conrad Jackson, David Rand, Kevin Lewis, Michael I. Norton, and Kurt Gray. 2017. Agent-based modeling: A guide for social psychologists. *Social Psychological and Personality Science* 8, 4 (2017), 387–395.
[67] C. Jones. 2020. The cerebral cortex realizes a universal probabilistic model of computation in complex Hilbert spaces. Preprint (2020).
[68] Ryan Julian, Benjamin Swanson, Gaurav S. Sukhatme, Sergey Levine, Chelsea Finn, and Karol Hausman. 2020. Never stop learning: The effectiveness of fine-tuning in robotic reinforcement learning. *arXiv preprint arXiv:2004.10190* (2020).
[69] A. Tversky and D. Kahneman. 1971. Belief in the law of small numbers. *Psychological Bulletin* 76, 2 (1971), 105–110.
[70] Tharindu Kaluarachchi, Andrew Reis, and Suranga Nanayakkara. 2021. A review of recent deep learning approaches in human-centered machine learning. *Sensors* 21, 7 (2021), 2514.

[71] Subbarao Kambhampati. 2019. Challenges of human-aware AI systems. *arXiv preprint arXiv:1910.07089* (2019).
[72] Matthew A. Kelly, Nipun Arora, Robert L. West, and David Reitter. 2019. High-dimensional vector spaces as the architecture of cognition. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci'19)*. 3491.
[73] William G. Kennedy. 2012. Modelling human behaviour in agent-based models. In *Agent-Based Models of Geographical Systems*. Springer, 167–179.
[74] Akif Quddus Khan, Salman Khan, and Utkurbek Safaev. 2022. Serious games and gamification: A systematic literature review. Virtual Reality & Intelligent Hardware 4, 3 (2022), 189–209.
[75] Irfan Khan, Xianchao Zhang, Mobashar Rehman, and Rahman Ali. 2020. A literature survey and empirical study of meta-learning for classifier selection. *IEEE Access* 8 (2020), 10262–10281.
[76] Jong Gwang Kim. 2021. Equilibrium computation of generalized Nash games: A new Lagrangian-based approach. *arXiv preprint arXiv:2106.00109* (2021).
[77] Suresh Kolekar, Shilpa Gite, Biswajeet Pradhan, and Ketan Kotecha. 2021. Behavior prediction of traffic actors for intelligent vehicle using artificial intelligence techniques: A review. *IEEE Access* 9 (2021), 135034–135058.
[78] Iuliia Kotseruba and John K. Tsotsos. 2020. 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review* 53, 1 (2020), 17–94.
[79] Max Kreminski, Ben Samuel, Edward Melcer, and Noah Wardrip-Fruin. 2019. Evaluating AI-based games through retellings. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 15. 45–51.
[80] Robert Tjarko Lange and Henning Sprekeler. 2020. Learning not to learn: Nature versus nurture in silico. *arXiv preprint arXiv:2010.04466* (2020).
[81] Christian Lebiere. 1999. The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft* 8, 1 (1999), 5–19.
[82] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.
[83] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).
[84] Falk Lieder and Thomas L. Griffiths. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* 43 (2020), 1–60.
[85] Falk Lieder, Paul M. Krueger, and Tom Griffiths. 2017. An automatic method for discovering rational heuristics for risky choice. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci'17)*.
[86] Antonio Lieto, Mehul Bhatt, Alessandro Oltramari, and David Vernon. 2018. The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research* 48 (2018), 1–3.
[87] Yang Liu, Yifeng Zeng, Yingke Chen, Jing Tang, and Yinghui Pan. 2019. Self-improving generative adversarial reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*. 52–60.
[88] Alexander Lotz, Nele Russwinkel, and Enrico Wohlfarth. 2020. Take-over expectation and criticality in level 3 automated driving: A test track study on take-over behavior in semi-trucks. *Cognition, Technology & Work* 22, 4 (2020), 733–744.
[89] Eric Mandelbaum, Isabel Won, Steven Gross, and Chaz Firestone. 2020. Can resources save rationality? "Anti-Bayesian" updating in cognition and perception. *Behavioral and Brain Sciences* 143 (2020), e16.
[90] Bruce G. Marcot and Trent D. Penman. 2019. Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental Modelling & Software* 111 (2019), 386–393.
[91] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. 2021. Core challenges of social robot navigation: A survey. *arXiv preprint arXiv:2103.05668* (2021).
[92] Smitha Milli, Falk Lieder, and Thomas L. Griffiths. 2021. A rational reinterpretation of dual-process theories. *Cognition* 217 (2021), 104881.
[93] Ronja Möller, Antonino Furnari, Sebastiano Battiato, Aki Härmä, and Giovanni Maria Farinella. 2021. A survey on human-aware robot navigation. *arXiv preprint arXiv:2106.11650* (2021).
[94] Catarina Moreira, Lauren Fell, Shahram Dehdashti, Peter Bruza, and Andreas Wichert. 2019. Towards a quantum-like cognitive architecture for decision-making. *arXiv preprint arXiv:1905.05176* (2019).
[95] Catarina Moreira and Andreas Wichert. 2018. Are quantum-like Bayesian networks more powerful than classical bayesian networks? *Journal of Mathematical Psychology* 82 (2018), 73–83.
[96] Cecily Morrison, Edward Cutrell, Martin Grayson, Anja Thieme, Alex Taylor, Geert Roumen, Camilla Longden, Sebastian Tschiatschek, Rita Faia Marques, and Abigail Sellen. 2021. Social sensemaking with AI: Designing an open-ended AI experience with a blind child. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[97] Finn Müller-Hansen, Maja Schlüter, Michael Mäs, Jonathan F. Donges, Jakob J. Kolb, Kirsten Thonicke, and Jobst Heitzig. 2017. Towards representing human behavior and decision making in earth system models—An overview of techniques and approaches. *Earth System Dynamics* 8, 4 (2017), 977–1007.
[98] Anis Najar and Mohamed Chetouani. 2021. Reinforcement learning with human advice: A survey. *Frontiers in Robotics and AI* 8 (2021), 584075.
[99] Neda Navidi. 2020. Human AI interaction loop training: New approach for interactive reinforcement learning. *arXiv preprint arXiv:2003.04203* (2020).
[100] Neda Navidi and Rene Landry. 2021. New approach in human-AI interaction by reinforcement-imitation learning. *Applied Sciences* 11, 7 (2021), 3068.
[101] Huansheng Ning, Rui Yin, Ata Ullah, and Feifei Shi. 2022. A survey on hybrid human-artificial intelligence for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2022), 6011–6026.
[102] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (2021), 48–62.
[103] Yael Niv. 2019. Learning task-state representations. *Nature Neuroscience* 22, 10 (2019), 1544–1553.
[104] Mike Oaksford and Nick Chater. 2001. The probabilistic approach to human reasoning. *Trends in Cognitive Sciences* 5, 8 (2001), 349–357.
[105] Afshin Oroojlooy, Mohammadreza Nazari, Davood Hajinezhad, and Jorge Silva. 2020. AttendLight: Universal attention-based reinforcement learning model for traffic signal control. *arXiv preprint arXiv:2010.05772* (2020).
[106] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters. 2018. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics* 7, 1-2 (2018), 1–179.
[107] Kei Ota, Devesh K. Jha, Diego Romeres, Jeroen van Baar, Kevin A. Smith, Takayuki Semitsu, Tomoaki Oiki, Alan Sullivan, Daniel Nikovski, and Joshua B. Tenenbaum. 2020. Towards human-level learning of complex physical puzzles. *arXiv E-prints CoRR abs/2011.07193* (2020).
[108] Kei Ota, Devesh K. Jha, Diego Romeres, Jeroen van Baar, Kevin A. Smith, Takayuki Semitsu, Tomoaki Oiki, Alan Sullivan, Daniel Nikovski, and Joshua B. Tenenbaum. 2021. Data-efficient learning for complex and real-time physical problem solving using augmented simulation. *IEEE Robotics and Automation Letters* 6, 2 (2021), 4241–4248.
[109] Alex Pentland and Andrew Liu. 1999. Modeling and prediction of human behavior. *Neural Computation* 11, 1 (1999), 229–242.
[110] Carlo Perrotta and Neil Selwyn. 2020. Deep learning goes to school: Toward a relational understanding of AI in education. *Learning, Media and Technology* 45, 3 (2020), 251–269.
[111] Cameron R. Peterson and Lee Roy Beach. 1967. Man as an intuitive statistician. *Psychological Bulletin* 68, 1 (1967), 29.
[112] Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. 2022. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *arXiv preprint arXiv:2203.01387* (2022).
[113] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. 2020. Watch-and-help: A challenge for social perception and human-AI collaboration. *arXiv preprint arXiv:2010.09890* (2020).
[114] Martin L. Puterman. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.
[115] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the International Conference on Machine Learning*. 4218–4227.
[116] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220* (2019).
[117] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486.
[118] Preeti Ramaraj, Charles L. Ortiz Jr., Matthew Klenk, and Shiwali Mohan. 2021. Unpacking human teachers' intentions for natural interactive task learning. *arXiv preprint arXiv:2102.06755* (2021).
[119] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2020. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *arXiv preprint arXiv:2010.07938* (2020).
[120] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. 2018. Shared autonomy via deep reinforcement learning. *arXiv preprint arXiv:1802.01744* (2018).
[121] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys* 54, 9 (2021), 1–40.
[122] Paul B. Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. 2014. Modeling human decision making in generalized Gaussian multiarmed bandits. *Proceedings of the IEEE* 102, 4 (2014), 544–571.
[123] Frank E. Ritter, Farnaz Tehranchi, and Jacob D. Oury. 2019. ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 10, 3 (2019), e1488.

[124] Nina Rizun and Yurii Taranenko. 2014. Simulation models of human decision-making processes. *Management Dynamics in the Knowledge Economy* 2, 2 (2014), 241–264.
[125] Stuart Russell and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall.
[126] Stuart Russell and Eric Wefald. 1991. Principles of metareasoning. *Artificial Intelligence* 49, 1-3 (1991), 361–395.
[127] Larry Samuelson. 1995. Bounded rationality and game theory. *Quarterly Review of Economics and Finance* 36 (1995), 17–36.
[128] Marlene Scharfe and Nele Russwinkel. 2019. A cognitive model for understanding the takeover in highly automated driving depending on the objective complexity of non-driving related tasks and the traffic environment. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci'19)*. 2734–2740.
[129] Marlene Scharfe and Nele Russwinkel. 2019. Towards a cognitive model of the takeover in highly automated driving for the improvement of human machine interaction. In *Proceedings of the 17th International Conference on Cognitive Modelling*.
[130] Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review* 21, 2 (2006), 97–126.
[131] Johannes Schneider. 2020. Humans learn too: Better human-AI interaction using optimized human inputs. *arXiv preprint arXiv:2009.09266* (2020).
[132] Francesco Semeraro, Alexander Griffiths, and Angelo Cangelosi. 2021. Human-robot collaboration and machine learning: A systematic review of recent research. *arXiv preprint arXiv:2110.07448* (2021).
[133] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.
[134] Guy Shani, Joelle Pineau, and Robert Kaplow. 2013. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems* 27, 1 (2013), 1–51.
[135] Herbert A. Simon. 1990. Bounded rationality. In *Utility and Probability*. Springer, 15–18.
[136] Meghendra Singh, Achla Marathe, Madhav V. Marathe, and Samarth Swarup. 2018. Behavior model calibration for epidemic simulations. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*. 1640–1648.
[137] Fabian H. Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S. Tolias. 2019. Engineering a less artificial intelligence. *Neuron* 103, 6 (2019), 967–979.
[138] Andrea Stocco, Catherine Sibert, Zoe Steine-Hanson, Natalie Koh, John E. Laird, Christian J. Lebiere, and Paul Rosenbloom. 2021. Analysis of the human connectome data supports the notion of a "Common Model of Cognition" for human and human-like intelligence across domains. *Neuroimage* 235 (2021), 118035.
[139] Alexander Streicher, Julius Busch, and Wolfgang Roller. 2021. Dynamic cognitive modeling for adaptive serious games. In *Proceedings of the International Conference on Human-Computer Interaction*. 167–184.
[140] Ron Sun and Sebastien Helie. 2012. Reasoning with heuristics and induction: An account based on the CLARION cognitive architecture. In *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN'12)*. IEEE, Los Alamitos, CA, 1–8.
[141] Jyrki Suomala. 2020. The consumer contextual decision-making model. *Frontiers in Psychology* 11 (2020), 2543.
[142] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press, Cambridge, MA.
[143] Alexandr Ten, Pramod Kaushik, Pierre-Yves Oudeyer, and Jacqueline Gottlieb. 2021. Humans monitor learning progress in curiosity-driven exploration. *Nature Communications* 12, 1 (2021), 1–10.
[144] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science* 331, 6022 (2011), 1279–1285.
[145] Robert Thomson, Christian Lebiere, John R. Anderson, and James Staszewski. 2015. A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition* 4, 3 (2015), 180–190.
[146] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. Mujoco: A physics engine for model-based control. In *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Los Alamitos, CA, 5026–5033.
[147] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.
[148] Tomer D. Ullman and Joshua B. Tenenbaum. 2020. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology* 2 (2020), 533–558.
[149] Christoph Urban and Bernd Schmidt. 2001. PECS–agent-based modelling of human behaviour. In *Emotional and Intelligent—The Tangled Knot of Social Cognition*. AAAI Fall Symposium Series. AAAI.
[150] Joaquin Vanschoren. 2019. Meta-learning. In *Automated Machine Learning*. Springer, Cham, Switzerland, 35–61.
[151] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 1–11.

[152] Jennifer Villareale and Jichen Zhu. 2021. Understanding mental models of AI through player-AI interaction. *arXiv preprint arXiv:2103.16168* (2021).
[153] Chen Wang, Pablo Cesar, and Erik Geelhoed. 2013. An invisible gorilla: Is it a matter of focus of attention? In *Proceedings of the Pacific-Rim Conference on Multimedia*. 318–326.
[154] Jane X. Wang. 2021. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences* 38 (2021), 90–95.
[155] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-AI interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.
[156] Rose E. Wang, J. Chase Kew, Dennis Lee, Tsang-Wei Edward Lee, Tingnan Zhang, Brian Ichter, Jie Tan, and Aleksandra Faust. 2020. Model-based reinforcement learning for decentralized multiagent rendezvous. *arXiv preprint arXiv:2003.06906* (2020).
[157] Rose E. Wang, Sarah A. Wu, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. 2020. Too many cooks: Coordinating multi-agent collaboration through inverse planning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*. 2032–2034.
[158] Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui Li. 2019. PressLight: Learning max pressure control to coordinate traffic signals in arterial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1290–1298.
[159] Jacob Whitehill. 2013. Understanding ACT-R—An outsider's perspective. *arXiv preprint arXiv:1306.0125* (2013).
[160] Uta Wilkens, Christian Cost Reyes, Tim Treude, and Annette Kluge. 2021. Understandings and perspectives of human-centered AI—A transdisciplinary literature review. *Frühjahrskongress der Gesellschaft für Arbeitswissenschaf, Bochum* (2021).
[161] Guojun Wu, Yanhua Li, Shikai Luo, Ge Song, Qichao Wang, Jing He, Jieping Ye, Xiaohu Qie, and Hongtu Zhu. 2020. A joint inverse reinforcement learning and deep learning model for drivers' behavioral prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. 2805–2812.
[162] Boming Xia, Xiaozhen Ye, and Adnan O. M. Abuassba. 2020. Recent research on AI in games. In *Proceedings of the 2020 International Wireless Communications and Mobile Computing Conference (IWCMC'20)*. IEEE, Los Alamitos, CA, 505–510.
[163] Annie Xie, Dylan P. Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. 2020. Learning latent representations to influence multi-agent interaction. *arXiv preprint arXiv:2011.06619* (2020).
[164] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems*. 1–13.
[165] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. 2020. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 193–202.
[166] Georgios N. Yannakakis and Julian Togelius. 2018. *Artificial Intelligence and Games*. Springer.
[167] R. Michael Young, Mark O. Riedl, Mark Branly, Arnav Jhala, R. J. Martin, and C. J. Saretto. 2004. An architecture for integrating plan-based behavior generation with interactive game environments. *Journal of Game Development* 1, 1 (2004), 1–29.
[168] Zahra Zahedi and Subbarao Kambhampati. 2021. Human-AI symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990* (2021).
[169] Yunqi Zhao, Igor Borovikov, Fernando de Mesentier Silva, Ahmad Beirami, Jason Rupert, Caedmon Somers, Jesse Harder, et al. 2020. Winning is not everything: Enhancing game development with intelligent agents. *IEEE Transactions on Games* 12, 2 (2020), 199–212.
[170] Boyuan Zheng, Sunny Verma, Jianlong Zhou, Ivor Tsang, and Fang Chen. 2021. Imitation learning: Progress, taxonomies and opportunities. *arXiv preprint arXiv:2106.12177* (2021).
[171] Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. 2020. Online Bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems* 33 (2020), 19238–19250.
[172] Jian-Qiao Zhu, Adam N. Sanborn, and Nick Chater. 2020. The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review* 127, 5 (2020), 719.
[173] Shlomo Zilberstein. 2011. Metareasoning and bounded rationality. In *Metareasoning: Thinking About Thinking*, Michael T. Cox and Anita Raja (Eds.). MIT Press, Cambridge, MA, 27–40.

Received 9 May 2022; revised 18 October 2022; accepted 9 January 2023