Current Robotics Reports (2020) 1:259–267
https://doi.org/10.1007/s43154-020-00019-0

A Survey of Mental Modeling Techniques in Human–Robot Teaming

Aaquib Tabrez[1] · Matthew B. Luebbers[1] · Bradley Hayes[1]

Published online: 31 August 2020

Abstract

Purpose of Review As robots become increasingly prevalent and capable, the complexity of roles and responsibilities assigned to them as well as our expectations for them will increase in kind. For these autonomous systems to operate safely and efficiently in human-populated environments, they will need to cooperate and coordinate with human teammates. Mental models provide a formal mechanism for achieving fluent and effective teamwork during human–robot interaction by enabling awareness between teammates and allowing for coordinated action.

Recent Findings Much recent research in human–robot interaction has made use of standardized and formalized mental modeling techniques to great effect, allowing for a wider breadth of scenarios in which a robotic agent can act as an effective and trustworthy teammate.

Summary This paper provides a structured overview of mental model theory and methodology as applied to human–robot teaming. Also discussed are evaluation methods and metrics for various aspects of mental modeling during human–robot interaction, as well as recent emerging applications and open challenges in the field.

Keywords Human-robot teaming · Mental models · Human-robot interaction · Theory of mind

Introduction

Traditionally, robots have worked separately from humans. Even in potentially collaborative environments like manufacturing, industrial robots most often operate in physically separated sections of the assembly floor. This work scheme of rigidly divided responsibility and prohibited human–robot interaction (HRI) prevails for reasons of safety and simplicity, but limits applications of these robots to strictly defined, well-structured, repetitive tasks [1]. Advances in autonomy are rapidly improving robots' ability to interact with, and even directly collaborate alongside, human teammates, opening up a wide range of new and impactful applications that leverage the unique skills of human and robot alike [2–5].

A key aspect of effective and fluent teamwork among humans is maintaining awareness of what teammates are likely to do or need, so as to coordinate actions. Humans tend to be adept at this task, and able to communicate plans and preferences easily understandable by their teammates [6]. Robots, however, do not have the benefit of human intuition. They must instead rely on explicit mathematical formalisms in order to approximate the mental states of human teammates and plan accordingly. This work focuses on characterizing recent work in developing these formalisms, known as *mental models*. In the following sections, we discuss the context and aims of mental model research for human–robot teaming, as well as describe and categorize the common methodologies, usage, and evaluation of such techniques.

This article belongs to the Topical Collection on *Service and Interactive Robotics*

✉ Aaquib Tabrez
mohd.tabrez@colorado.edu

Matthew B. Luebbers
matthew.luebbers@colorado.edu

Bradley Hayes
bradley.hayes@colorado.edu

[1] Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309, USA

Mental Models

Mental models, also referred to as *mental representations* in psychology, are organized knowledge structures that allow individuals to interact with their environment [7]. Although the mental model has been used as an explanatory mechanism in a variety of disciplines over the years, its root can be traced back to twentieth century psychology and epistemology. In 1943, Kenneth Craik posited in his seminal work that the mind provides a "small-scale model" of reality, enabling us to predict events [8]. In essence, mental models serve the crucial purpose of helping people to describe, explain, and predict events in their environment [9]. Since then, mental models have gained popularity in the human factors community for their effectiveness in eliciting and strengthening teamwork fluency for complex task execution, such as in tactical military operations [10, 11]. Inspired by this success, several architectures for HRI have since replicated this fluency and teamwork by developing mental modeling techniques for robotic agents that operate in human-populated environments.

In HRI literature, the concept of mental modeling is often conflated or used interchangeably with another important concept in developmental psychology: *Theory of Mind* (ToM). To be capable of ToM simply denotes an ability to attribute thought, desires, and intentions to others [12]. Theory of Mind is crucial for everyday human social interactions (e.g., for analyzing, judging, and inferring others' behaviors), with evidence that typically developing humans exhibit this capability by the age of 5 [13]. Accordingly, several architectures for human–robot teaming in HRI incorporate aspects of a ToM for other agents [14–19].

In general, mental models and ToM go hand in hand during human–robot interaction, as a robot modeling other agents is analogous to having an agent with a ToM capacity. Furthermore, it leads to an interesting phenomenon during human–robot teaming as humans also form a ToM directed at their robot teammate. Therefore, mental modeling enables a phenomenon where a robot may form a belief over a human's mental model of the robot. This meta-modeling is defined as second-order mental modeling which enables robots to estimate how a human's mental model is affected by its own behavior [20]. Thus, current work in mental modeling for human–robot teaming can be broadly classified into first-order (or standard) or second-order mental models.

We can see how effective mental models correlate with team functioning: team members predict what their teammates will do or need, facilitating the coordination of actions. Prior studies in the human factors community demonstrate a positive relationship between team performance and similarity between the mental models of team members [9, 21, 22]. This implies that shared understanding of the team is a crucial factor of effective team performance (i.e., team members should have a shared mental model). Shared mental model (SMM) theory states that team members should hold compatible mental models that lead to common expectations for shared task execution to avoid failure [23, 24]. To summarize, if a mental model helps in describing, explaining, and predicting the behavior of a system, a shared mental model serves the purpose of describing, explaining, and predicting the behavior of a team.

Mental Models in Human–Robot Teaming

Teamwork is the collaborative effect of a group's effort toward achieving a common goal [25]. In the mental modeling literature, collaborative tasks are often broken up into smaller submodels representing components of effective teamwork, such as models of task procedures and strategies, models of inter-member interaction and information flow, or models of individual team member skill and preferences [9].

These various types of mental models and their incorporation of shared knowledge in teams help in achieving characteristic traits such as fluent behavior between teammates, quick adaptation to changing task demands, trusting collaborators with roles and responsibilities, effective communication, and decision making in time-critical applications. Several studies in human-robot collaboration have attempted to elicit these positive qualities through the use of mental models. In this section, we present a systematic characterization of desirable traits which can be achieved through mental modeling in human–robot teaming:

- **Fluent behavior**: Fluency, as defined by Hoffman, is a "coordinated meshing of joint activities between members of a well-synchronized team" [26•]. This quality of interaction, collaborative fluency, intuitively means human and robot are well synchronized in timing, and they can alter plans and actions appropriately, and often without much communication.
- **Adaptability**: During collaboration, plans change, and team members (both human and robot) should be able to alter their plans and actions appropriately and dynamically as needed. Previous studies show that shared or common mental models can be leveraged for changing task demands for quick adaptation in a team [23, 27•].
- **Trust building**: Trust is a critical element for the success of a team. In human–robot interaction, studies show that people trust a collaborative robot when they can discern its role and responsibility, have confidence

in its capabilities, and possess an accurate understanding of its decision-making process (a shared mental model) [28, 29].

- **Effective communication**: Information exchange, either verbal or non verbal, is pivotal for collaboration. A collaborative agent can leverage mental models to warn its human teammate about potential failures or ask for help when it is unable to complete a task [30, 31].
- **Explainability**: Knowledge sharing and expectation matching also have importance for behavior explainability [32–34]. The recent surge in popularity of explainable AI (xAI) has shown the crucial importance of agents' ability to explain their decision-making process, leading to improved transparency, trust, and team performance.

Mental Model Methodologies

In this section, we discuss successful methods for mental modeling in human-robot teaming contexts. We organize the literature into three categories: first-order (or standard) mental models, second-order mental models, and shared mental models.

First-Order Mental Models

In first-order mental models, robots model the behavior of human collaborators to infer their beliefs, intentions, and goals, for the purpose of predicting their actions. Usually, such modeling can be functionally broken down into two steps which a framework must resolve: (1) the human's reward function (which motivates the human's behavior in the world), and (2) a planning algorithm which connects that inferred reward function to robot behavior [35].

One of the simplest approaches is based on the principle of rationality [36, 37]: the expectation that agents will plan approximately rationally to achieve their goals, given their beliefs about the world (i.e., they will take actions that maximize their expected reward). One way to infer a human's reward function is to observe their behavior through inverse reinforcement learning (IRL). For example, the widely used maximum entropy IRL formulation optimizes a model to fit a reward function that incentivizes a human demonstrator's actions exponentially more than unobserved actions [38, 39].

A similar approach to inferring a human's reward function is through inverse planning. Baker et al. propose a computational framework based on Bayesian inverse planning for modeling human action understanding. They modeled human decision making as rational probabilistic planning with Markov decision processes (MDPs), and inverted this relation using Bayes' rule to infer agents' beliefs and goals from their actions (running the principle of rationality in reverse) [40, 41]. They were able to extend this method to a Bayesian model of Theory of Mind (BToM), which provides the predictive model of belief and desire-dependent action (the ToM capacity of the collaborative human) as a Partially Observable Markov Decision Process (POMDP) [42], and reconstructs an agent's joint belief state and reward function using Bayesian inference based on observations of the agent's behavior [43, 44].

From a planning and decision-making point of view, the noisy rational choice model (also known as Boltzmann rational) [45, 46] is a popular method in robotics where actions or trajectories are chosen in proportion to their exponentiated reward. Here, it is assumed that the collaborative agent has access to some underlying human reward function (usually inferred through IRL or inverse planning approaches). The human is modeled to act rationally with the highest probability, but with a non-zero probability of behaving sub-optimally [20, 47–50].

Humans frequently deviate from rational behavior due to specific biases such as time pressures, loss aversion, and the like [51]. Furthermore, they are limited in cognitive capacity, which leads to forgetfulness, limited planning horizons, and false beliefs. Some recent methods attempt to introduce these inconsistencies to the rational model assumption [52]. Nikolaidis et al. gave a Bounded-Memory Adaptation Model, which models humans as boundedly rational and subject to memory and recency constraints, through a probabilistic finite-state controller that captures human adaptive behaviors [19]. Kwon et al. used a risk-aware human model from behavioral economics (Cumulative Prospect Theory) for modeling loss aversion behaviors of humans under risk and uncertainty [53].

Another recent approach for human behavior modeling is the Reward Augmentation and Repair through Explanation (RARE) framework for estimating and improving a collaborators' task understanding. Here, Tabrez et al. provided a computational framework for human reward function estimation via a set of possible Hidden Markov Models (HMMs) [30], representing a task's reward function and partially deficient variants (e.g., missing reward information). The collaborative agent must infer the most likely HMM for explaining the teammates' behavior, which in turn indicates a plausible underlying reward function for explaining the human's actions.

Second-Order Mental Models

The concept of a second-order mental model is related to a recursive type of reasoning modeled by game theorists ("I believe that you believe that I believe...") which can be extended to a possibly infinite reasoning process [54, 55]. The second-order mental model is one step deeper in behavior modeling (i.e., a robot forming a belief over a human's model of the robot). Second-order mental models enable robots to possess more predictable and explicable behavior, as the effects of their actions on another agent's perception of them are included in the model.

Work by Huang et al. modeled humans as learning a robot's objective function over time by observing its behavior using Bayesian IRL, an inversion of typical IRL paradigms where a robotic agent attempts to infer human objective functions. To account for noisy learning behavior from humans, the authors utilize approximate-inference models. Using this insight, an agent can plan for actions that communicate to the human so as to be maximally informative, better enabling humans to anticipate what the robot will do in novel situations [56].

Another approach that has shown promise is the Interactive POMDP (I-POMDP) framework, which modifies a traditional single-agent POMDP to include other agents by creating the notion of an interactive state. An interactive state encapsulates both the environment state and the modeled belief state attributed to another agent. Brooks and Szafir use this I-POMDP framework [57] for performing Bayesian inference of second-order mental models. They estimate the human's Q-function (a function that helps determine the optimal action given an interactive state) through IRL and use it to infer the human's belief state about the agent, by comparing it with the human's actions assuming a Boltzmann rational behavior model [20].

Shared Mental Models

Shared mental models enable team members to draw on their own well-structured common knowledge as a basis for selecting actions that are consistent and coordinated with those of their teammates. They are strongly correlated to team performance [9]. In this section, we focus on methods employed for establishing a shared understanding between teammates.

One well-known approach in HRI inspired by SMM is work on human–robot cross-training by Nikolaidis and Shah, which focuses on computing a robot policy aligned with human preference by iteratively switching roles (between a human and a robot) to learn a shared plan for a collaborative task [58]. Hadfield-Menell et al. approached SMM as a value alignment problem, ensuring that the agents behave in alignment with human values. They utilize a cooperative inverse reinforcement learning (CIRL) formulation, where a robot maximizes a human teammate's unknown reward in a cooperative, partial information game. They show that solutions within this formalism result in active teaching and active learning behaviors [59].

Nikolaidis et al. also propose a game-theoretic model of a human's partial adaptation to a robot teammate. This method assumes the robot agent knows a "true" utility function for the team, and the human is following a best-response strategy to the robot action based on their own, possibly incorrect reward function. The robot uses this model to decide optimally between revealing information to the human and choosing the best action given the information that the human currently has [27•].

From these well-known models, we can see that establishing a shared mental model requires communication between agents (except the cross-training method, where agents learn each other's responsibilities by switching roles). We can separate these communication strategies into two categories: implicit (e.g., using movement or motion) and explicit (e.g., verbal explanations).

Implicit communicative models. A popular principle in motion planning for expressing intention to a collaborator is the notion of legibility. Dragan et al. developed a formalism to mathematically define and distinguish predictability (predicting a trajectory given a known goal) and legibility (predicting a goal given an observed trajectory) of motion based on a rational action assumption for the collaborative human [50]. Kulkarni et al. generate explicable robot behavior by learning a regression model over plan distances and mapping them to a labeling scheme used by a human observer, minimizing divergence between the robot's plan and the plan expected by the human [60].

Another mode of implicit communication is through gesture and non-verbal expression. One example of this is work by Lee et al. which uses a BToM approach to model dyadic storytelling interactions [61]. They propose a method for a robot to influence and infer the mental state of a child while telling it a story, specifically estimating the child's degree of attentiveness toward the robot. They model emotion expression as a joint process of estimating people's beliefs through inference inversion using a Dynamic Bayesian Network (DBN), and subsequently produce nonverbal expressions (speaker cues) to affect those beliefs (attention state).

Explicit communicative models. Model reconciliation processes try to identify and resolve the model differences of a collaborator through explanations, thereby establishing a shared mental model. These processes lead to predictable behavior from the collaborative agent: a consequence of explainability [62–64]. Briggs and Scheutz's recent work provides a formal framework to correct false or missing beliefs of collaborators in a transparent and human-like manner by using adverbial cues, adhering to Grice's maxims [65] of effective conversational communication (quality, quantity, and relevance) [66]. Additional recent works also address the generation of these explanations, seeking output that is optimal with respect to various quantitative and qualitative criteria including selectivity, contrastiveness, and succinctness [29, 67–69•].

Evaluation Methods

In this section, we discuss evaluation methods employed in human-robot teaming for each of the desirable traits characterised in Section "Mental Models."

Team Fluency Fluency, the metric for well-synchronized meshing of joint actions between humans and robots, is difficult to measure and optimize in practice [70]. Hoffman and Breazeal demonstrated that fluency is a distinct construct to efficiency through a user study involving an anticipatory controller (when the robot anticipated participants' actions, task efficiency was not improved, but participants' sense of fluency was increased) [71]. For team fluency, there exist a number of validated subjective metric scales, as well as commonly used objective measures, such as human and robot idle time, fraction of time spent concurrently working between agents, and delay times between one agent finishing a precursor task and another agent resuming that task [26•].

Adaptability Shared mental models offer a mechanism for adaptability: quick, on the fly strategy adjustments by a team. As adaptability is intrinsically linked to performance, a majority of measures are objective, often treating an adaptable controller as an independent variable to compare alongside other controllers. Specific objective measures vary with the formulation used, including mean reward accrued [27•] and similarity metrics between human and robot notions of "correct action sequence" in an evolving task [58]. Though there is a notable lack of validated subjective measures for agent adaptability in HRI, many studies utilize subjective metric scales for correlated measures such as team fluency and trustworthiness [26•, 58]. Nikolaidis et al. have additionally showed that accounting for individual differences in humans' willingness to adapt to a robot is positively correlated with trust [19].

Team Trust Shared mental models promote trust and reliability by alleviating uncertainty in roles, responsibilities, and capabilities while working in a team. Lee and See proposed a three-dimensional model wherein trust is influenced by a person's knowledge of what the robot is supposed to do (purpose), how it functions (process), and its performance [72]. Based on previous studies, robot performance is considered to be the most influential factor for trust [73], likely due to the importance of the agent's ability to meet expectations [74]. Other factors with positive relationships to trust are minimizing system fault occurrence, system predictability, and transparency [75]. Most subjective measures for trust in HRI research are newly created to match individual study requirements and lack the rigor in development and validation available in standardized scales from the human factors community. Some well-known standardized scales with high potential for use in HRI to evaluate a user's trust perception of an agent are the HRI Trust Scale, Dyadic Trust Scale (DTS), and Robotic Social Attributes Scale (RoSAS) [75, 76].

Effective Communication Previous studies show that information exchange and effective communication are important for building trust between team members. These communications can be verbal (explicit) or nonverbal (implicit), as seen in Section "Mental Model Methodologies." For explicit models, the following qualities have been found to be positively correlated with trust and teamwork: task-related communications, contrastive explanations expressing model divergence, and user and context-dependent information (such as providing technical information to an expert, and accessible information to a lay-user) [77–79]. For implicit models, such as those aimed at plan legibility and explicability, self-reported understanding of a robotic agents' behavior or goal is a common evaluation metric. Additionally, subjective metrics are often crafted for individual study requirements, aimed at uncovering related traits like robot trustworthiness [50, 80, 81].

Explainability Explainability deals with the understanding of the mechanisms by which a robot operates and the ability to explain robots' behavior or underlying logic [30, 68]. Existing works in explainable AI assess the effects of explainability through self-reported understanding of the agent behavior, successful task completions, system faults, task completion time, number of irreparable mistakes, and trust in automation. A survey by Walkotter et al. described three categories of measures for evaluating the effectiveness of explainable architectures (in descending order of importance): (1) trust (willingness of users to agree with robot decisions through a self-reported scale); (2) robustness (failure avoidance during the interaction); and (3) efficiency (how quickly tasks are completed) [82].

Emerging Fields and Discussion

Mental models have proven beneficial for many human-robot teaming applications such as assistive and healthcare robotics [4], social path planning and navigation [5], search and rescue [2], and autonomous driving [53, 83]. In this section, we describe a selection of more recent emerging use cases of mental models in HRI.

Though robots have been fixtures in industrial applications since the 1970s [84], the factory of the future is likely to utilize robots for a much broader range of tasks, and in a much more collaborative manner, enabled in part through the use of recent developments in mental models.

Many of these potential robot tasks intrinsically require operation in proximity to humans, raising issues of safety and efficiency. Recent work by Unhelkar et al. provides a framework for human-aware task and motion planning in shared-environment manufacturing [85]. Additional research in this area focuses on the problem of task scheduling for safely and effectively coordinating human and robot agents in resource-constrained environments [86, 87]. Another recent development has been toward the generation of supporting behavior for improving human collaborators' task performance. These supportive behaviors do not directly contribute to a task but instead alleviate the cognitive and kinematic burdens of a collaborating human (e.g., fetching tools or stabilizing objects during assembly) [62, 88].

Furthermore, developments in augmented reality (AR) technology have shown promise for industrial HRI applications. AR represents a novel modality of model communication for human–robot collaboration, wherein details of a robot's plan or decision making process are visualized and presented to a human teammate as holographic imagery overlaid onto the robot itself, viewed through a head-mounted display. Notable work in this area has focused on visually conveying robotic motion intent during human-robot teaming tasks with AR, both for robotic manufacturing arms [89], and mobile robots [90], a technique which has been shown to broadly increase objective measures of task accuracy and efficiency, as well as subjective perceptions of robot transparency and trustworthiness. Recent work has explored the inclusion of human-to-robot communication features on top of AR visualization, allowing human teammates to diagnose problems with and modify a robot's plans or internal models during collaboration [91, 92].

Behavior manipulation, also known as policy elicitation, refers to a class of problems in human-robot teaming wherein an agent must guide humans toward an optimal policy (or away from potential failure states) in order to successfully complete a task, either through implicit or explicit communication [30, 69•, 93]. Consider an emergency evacuation scenario, where an agent is tasked with guiding people safely out of a building. The agent could steer evacuees away from a possible hazardous state either by blocking their path or by verbally updating their internal model ("fire in next hallway") to encourage alternative, less dangerous paths. Various challenges related to behavior manipulation include accurately modeling human behavior [30], leveraging human models to find failure modes [94], and succinctly generating persuasive human intelligible semantic updates (or executing mitigating actions) [68]. This concept of behavior modeling has additionally been extended to intelligent teaching or coaching for effective personalized learning [95].

With the currently observed rate of increase in agents' capability for social behavior and natural language generation, important problems surface regarding robot ethics and norms [96, 97], particularly in cases of policy elicitation (manipulating the human in the hopes of achieving some greater good). These behaviors and capabilities induce perceptions of a moral and social agency in robots similar to human standards of morality [98]. In reality, such actions/behaviors do not embody any maliciousness but rather emerge due to necessity of situation and cooperation. Some major challenges within this domain of problems include establishing moral norms during collaboration, anticipating possible norm violations, attempting to prevent them while executing, and if norms are eventually violated, taking mitigating actions to create transparency and user awareness (such as providing justifiable explanations communicating the robot's decision-making processes or capabilities) [99, 100].

As evidenced by the emerging application areas found within human–robot teaming literature, mental models continue to be developed and applied in novel ways. Research in human–robot interaction is rapidly evolving and expanding into new application areas, so this list is far from exhaustive. In this survey, we have provided a general overview of mental models as applied to human–robot teaming: formalisms which have proven to be significantly beneficial for fluent collaboration and cooperation between teammates. As evident in this summary, there are many exciting developments within this space, as well as many open and challenging problems to drive future research.

Funding Information This work was funded in part by NSF Award #1830686 and the U.S. Army Research Lab STRONG program.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:

- • Of importance
- •• Of major importance

1. Engelberger J. F. Robotics in practice: management and applications of industrial robots Springer Science & Business Media. 2012.

2. Nourbakhsh I. R., Sycara K., Koes M., Yong M., Lewis M., Burion S. Human-robot teaming for search and rescue. IEEE Pervasive Computing. 2005;4(1):72–79.
3. Nikolaidis S, Lasota P., Rossano G., Martinez C., Fuhlbrigge T., Shah J. Human-robot collaboration in manufacturing: Quantitative evaluation of predictable, convergent joint action. In: IEEE ISR 2013, pp. 1–6 IEEE; 2013.
4. Dragan A. D., Srinivasa S. S. Formalizing assistive teleoperation MIT Press. 2012.
5. Rios-Martinez J., Spalanzani A., Laugier C. From proxemics theory to socially-aware navigation: a survey. Int. J. Soc. Robot. 2015;7(2):137–153.
6. Cohen P. R., Levesque H. J., Nunes J. H., Oviatt S. L. Task-oriented dialogue as a consequence of joint activity. In: Proceedings of PRICAI-90; 1990. p. 203–208.
7. Wilson J. R., Rutherford A. Mental models: Theory and application in human factors. Hum. Factors. 1989;31(6):617–634.
8. Craik K. J. W. The nature of explanation, vol. 445 CUP Archive. 1952.
9. Mathieu J. E., Heffner T. S., Goodwin G. F., Salas E., Cannon-Bowers J. A. The influence of shared mental models on team process and performance. Journal of applied psychology. 2000;85(2):273.
10. Cooke N. J., Salas E., Cannon-Bowers J. A., Stout R. J. Measuring team knowledge. Human factors. 2000;42(1):151–173.
11. Marks M. A., Zaccaro S. J., Mathieu J. E. Performance implications of leader briefings and team-interaction training for team adaptation to novel environments. Journal of applied psychology. 2000;85(6):971.
12. Premack D., Woodruff G. Does the chimpanzee have a theory of mind?. Behavioral and brain sciences. 1978;1(4):515–526.
13. Gopnik A., Sobel D. M., Schulz L. E., Glymour C. Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. Developmental psychology. 2001;37(5): 620.
14. Devin S., Alami R. An implemented theory of mind to improve human-robot shared plans execution. 2016.
15. Zhao Y., Holtzen S., Gao T., Zhu S.-C. Represent and infer human theory of mind for human-robot interaction. In: 2015 AAAI fall symposium series, vol. 2; 2015.
16. Görür O. C., Rosman B. S., Hoffman G., Albayrak S. Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention. 2017.
17. Scassellati B. Theory of mind for a humanoid robot. Auton. Robot. 2002;12(1):13–24.
18. Leyzberg D, Spaulding S., Scassellati B. Personalizing robot tutors to individuals' learning differences. In: Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction, pp. 423–430 ACM; 2014.
19. Nikolaidis S, Zhu Y. X., Hsu D., Srinivasa S. Human-robot mutual adaptation in shared autonomy. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction pp. 294–302 ACM; 2017.
20. Brooks C., Szafir D. Building second-order mental models for human-robot interaction. 2019. arXiv:1909.06508.
21. Bolstad C. A., Endsley M. R. Shared mental models and shared displays: an empirical evaluation of team performance. In: proceedings of the human factors and ergonomics society annual meeting vol. 43, pp. 213–217, SAGE Publications Sage CA: Los Angeles CA; 1999.
22. Minsky M. A framework for representing knowledge. 1974.
23. Converse S., Cannon-Bowers J., Salas E. Shared mental models in expert team decision making. Individual and group decision making: Current issues. 1993;221:221–46.
24. Jonker C. M., van Riemsdijk M. B., Vermeulen B. Shared mental models. In: Coordination, Organizations, Institutions, and Norms in Agent Systems VI M. De Vos, N. Fornara, J.V. Pitt, and G. Vouros, eds. Berlin, Heidelberg, pp. 132–151 Springer Berlin Heidelberg; 2011.
25. Salas E., Cooke N. J., Rosen M. A. On teams, teamwork, and team performance: discoveries and developments. Human factors. 2008;50(3):540–547.
26.• Hoffman G. Evaluating fluency in human–robot collaboration. IEEE Transactions on Human-Machine Systems. 2019;49(3):209–218. **This work provides a thorough description of human–robot teaming fluency along with several metrics for evaluating fluency in human–robot shared-location teamwork.**
27.• Nikolaidis S., Nath S., Procaccia A. D., Srinivasa S. Game-theoretic modeling of human adaptation in human-robot collaboration. In: Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction pp. 323–331; 2017. **This work presents a study showing that expectation matching between collaborators (i.e., a robot revealing its capabilities and intent) leads to significantly improved human–robot team performance**.
28. Arnold M., Bellamy R. K., Hind M., Houde S., Mehta S., Mojsilović A., Nair R., Ramamurthy K. N., Olteanu A., Piorkowski D., et al. Factsheets: increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, vol. 63, no. 2019;4(5):6–1.
29. Tabrez A, Hayes B. Improving human-robot interaction through explainable reinforcement learning. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) pp. 751–753 IEEE; 2019.
30. Tabrez A, Agrawal S., Hayes B. Explanation-based reward coaching to improve human performance via reinforcement learning. 2019.
31. Tellex S., Knepper R., Li V, Rus D., Roy N. Asking for help using inverse semantics. 2014.
32. Wang N., Pynadath D. V., Hill S. G. Trust calibration within a human-robot team: comparing automatically generated explanations. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction pp. 109–116 IEEE Press; 2016.
33. Miller T. Explanation in artificial intelligence: insights from the social sciences Artificial Intelligence. 2018.
34. Viganò L., Magazzeni D. Explainable security. 2018. arXiv:1807.04178.
35. Armstrong S., Mindermann S. Occam's razor is insufficient to infer the preferences of irrational agents. 2018.
36. Dennett D. C. The intentional stance MIT press. 1989.
37. Gergely G., Nádasdy Z., Csibra G., Bíró S. Taking the intentional stance at 12 months of age. Cognition. 1995;56(2):165–193.
38. Ng A. Y., Russell S. J., et al. Algorithms for inverse reinforcement learning. In: Icml; 2000. p. 2.
39. Ziebart B. D., Maas A. L., Bagnell J. A., Dey A. K. Maximum entropy inverse reinforcement learning. In: Aaai, vol. 8, pp. 1433–1438 Chicago IL USA; 2008.
40. Baker C. L., Saxe R., Tenenbaum J. B. Action understanding as inverse planning. Cognition. 2009;113(3):329–349.
41. Baker C. L., Tenenbaum J. B., Saxe R. R. Goal inference as inverse planning. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 29; 2007.

42. Kaelbling L. P., Littman M. L., Cassandra A. R. Planning and acting in partially observable stochastic domains. Artificial intelligence. 1998;101(1-2):99–134.
43. Baker C., Saxe R., Tenenbaum J. Bayesian theory of mind: modeling joint belief-desire attribution. In: Proceedings of the annual meeting of the cognitive science society, vol. 33; 2011.
44. Baker C. L., Tenenbaum J. B. Modeling human plan recognition using bayesian theory of mind. 2014.
45. Otsuka M., Osogami T. A deep choice model. In: Thirtieth AAAI Conference on Artificial Intelligence; 2016.
46. Osogami T, Otsuka M. Restricted Boltzmann machines modeling human choice. In: Advances in Neural Information Processing Systems; 2014. p. 73–81.
47. Sadigh D., Dragan A. D., Sastry S., Seshia S. A. Active preference-based learning of reward functions. In: robotics: Science and Systems; 2017.
48. Pellegrinelli S, Admoni H., Javdani S., Srinivasa S. Human-robot shared workspace collaboration via hindsight optimization. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 831–838 IEEE; 2016.
49. Palan M., Landolfi N. C., Shevchuk G., Sadigh D. Learning reward functions by integrating human demonstrations and preferences. 2019. arXiv:1906.08928.
50. Dragan A. D., Lee K. C., Srinivasa S. S. Legibility and predictability of robot motion. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 301–308 IEEE; 2013.
51. Tversky A., Kahneman D. Judgment under uncertainty:, heuristics and biases. science. 1974;185(4157):1124–1131.
52. Simon H. A. Rational decision making in business organizations. The American economic review. 1979;69(4):493–513.
53. Kwon M., Biyik E., Talati A., Bhasin K., Losey D. P., Sadigh D. 2020. arXiv:2001.04377.
54. Gmytrasiewicz P. J., Durfee E. H. Rational coordination in multi-agent environments. Auton. Agent. Multi-Agent Syst. 2000;3(4):319–350.
55. Wilks Y., Ballim A. Multiple agents and the heuristic ascription of belief Computing Research Laboratory New Mexico State University. 1986.
56. Huang S. H., Held D., Abbeel P., Dragan A. D. Enabling robots to communicate their objectives. Auton. Robot. 2019;43(2):309–326.
57. Gmytrasiewicz P. J., Doshi P. A framework for sequential planning in multi-agent settings. J. Artif. Intell. Res. 2005;24:49–79.
58. Nikolaidis S, Shah J. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI) pp. 33–40 IEEE; 2013.
59. Hadfield-Menell D., Russell S. J., Abbeel P., Dragan A. Cooperative inverse reinforcement learning. In: Advances in neural information processing systems; 2016. p. 3909–3917.
60. Kulkarni A., Zha Y., Chakraborti T., Vadlamudi S. G., Zhang Y., Kambhampati S. Explicablility as minimizing distance from expected behavior. 2016. arXiv:1611.05497.
61. Lee J. J., Sha F., Breazeal C. A Bayesian theory of mind approach to nonverbal communication. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) pp. 487–496 IEEE; 2019.
62. Hayes B., Scassellati B. Effective robot teammate behaviors for supporting sequential manipulation tasks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2015.
63. Chakraborti T., Kambhampati S., Scheutz M., Zhang Y. AI challenges in human-robot cognitive teaming. 2017. arXiv:1707.04775.
64. Dragan A. D. 2017. arXiv:1705.04226.
65. Grice H. P. Logic and conversation. In: Speech acts, pp. 41–58 Brill; 1975.
66. Briggs G., Scheutz M. Facilitating mental modeling in collaborative human-robot interaction through adverbial cues. In: proceedings of the SIGDIAL 2011 Conference; 2011. p. 239–247.
67. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. 2019;267:1–38.
68. Hayes B., Shah J. A. Improving robot controller transparency through autonomous policy explanation. In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction HRI pp. 303–312 IEEE; 2017.
69.• Chakraborti T., Sreedharan S., Zhang Y., Kambhampati S. Plan explanations as model reconciliation: moving beyond explanation as soliloquy. 2017. **This work characterizes the problem of model reconciliation, wherein a AI system suggests changes to a human teammate's model through explanation based on the divergence between their respective models.**
70. Thomaz A., Hoffman G., Cakmak M. Computational human-robot interaction. Foundations and Trends in Robotics. 2016;4(2-3):105–223.
71. Hoffman G., Breazeal C. Cost-based anticipatory action selection for human–robot fluency. IEEE transactions on robotics. 2007;23(5):952–961.
72. Lee J. D., See K. A. Trust in automation: designing for appropriate reliance. Human factors. 2004;46(1):50–80.
73. Hancock P. A., Billings D. R., Schaefer K. E., Chen J. Y., De Visser E. J., Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. Human factors. 2011;53(5):517–527.
74. Kwon M., Jung M. F., Knepper R. A. Human expectations of social robots. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 463–464 IEEE; 2016.
75. Lewis M., Sycara K., Walker P. The role of trust in human-robot interaction, pp. 135–159 cham: Springer International Publishing. 2018.
76. Sebo S. S., Krishnamurthi P., Scassellati B. I don't believe you : investigating the effects of robot trust violation and repair. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) pp. 57–65 IEEE; 2019.
77. Zahedi Z, Olmo A., Chakraborti T., Sreedharan S., Kambhampati S. Towards understanding user preferences for explanation types in model reconciliation,. 2019.
78. Ciocirlan S.-D., Agrigoroaie R., Tapus A. Human-robot team: effects of communication in analyzing trust. 2019.
79. Chakraborti T., Sreedharan S., Grover S., Kambhampati S. Plan explanations as model reconciliation–an empirical study. 2018. arXiv:1802.01013.
80. Kwon M., Huang S. H., Dragan A. D. Expressing robot incapability. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction; 2018. p. 87–95.
81. Kulkarni A., Zha Y., Chakraborti T., Vadlamudi S. G., Zhang Y., Kambhampati S. Explicable planning as minimizing distance from expected behavior. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems pp. 2075–2077 International Foundation for Autonomous Agents and Multiagent Systems; 2019.

82. Wallkotter S., Tulli S., Castellano G., Paiva A., Chetouani M. Explainable agents through social cues: a review. 2020. arXiv:2003.05251.
83. Sadigh D., Landolfi N., Sastry S. S., Seshia S. A., Dragan A. D. Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state. Auton. Robot. 2018;42(7):1405–1426.
84. Hägele M., Nilsson K., Pires J. N., Bischoff R. Industrial robotics. In: Springer handbook of robotics, pp. 1385–1422 Springer; 2016.
85. Unhelkar V. V., Lasota P. A., Tyroller Q., Buhai R.-D., Marceau L., Deml B., Shah J. A. Human-aware robotic assistant for collaborative assembly: integrating human motion prediction with planning in time. IEEE Robotics and Automation Letters. 2018;3(3):2394–2401.
86. Chakraborti T., Zhang Y., Smith D. E., Kambhampati S. Planning with resource conflicts in human-robot cohabitation. In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems; 2016. p. 1069–1077.
87. Gombolay M., Wilcox R., Shah J. Fast scheduling of multi-robot teams with temporospatial constraints. 2013.
88. Baraglia J, Cakmak M., Nagai Y., Rao R., Asada M. Initiative in robot assistance during collaborative task execution. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI) pp. 67–74 IEEE; 2016.
89. Rosen E., Whitney D., Phillips E., Chien G., Tompkin J., Konidaris G., Tellex S. Communicating robot arm motion intent through mixed reality head-mounted displays. In: Robotics Research pp. 301–316 Springer; 2020.
90. Walker M., Hedayati H., Lee J., Szafir D. Communicating robot motion intent with augmented reality. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction; 2018. p. 316–324.
91. Luebbers M. B., Brooks C., Kim M. J., Szafir D., Hayes B. Augmented reality interface for constrained learning from demonstration. In: Proceedings of the 2nd International Workshop on Virtual, Augmented and Mixed Reality for HRI (VAM-HRI); 2019.
92. Gregory J. M., Reardon C., Lee K., White G., Ng K., Sims C. Enabling intuitive human-robot teaming using augmented reality and gesture control. In: arXiv:1909.06415; 2019.
93. Sadigh D., Sastry S., Seshia S. A., Dragan A. D. Planning for autonomous cars that leverage effects on human actions. In: robotics: Science and Systems 2 Ann Arbor, MI USA; 2016.
94. Tabrez A., Luebbers M. B., Hayes B. Automated failure-mode clustering and labeling for informed car-to-driver handover in autonomous vehicles; 2020. arXiv:2005.04439.
95. Leyzberg D., Ramachandran A., Scassellati B. The effect of personalization in longer-term robot tutoring. ACM Transactions on Human-Robot Interaction (THRI). 2018;7(3):19.
96. Williams T., Zhu Q., Wen R., de Visser E. J. The Confucian matador: three defenses against the mechanical bull. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction; 2020. p. 25–33.
97. Jackson R. B., Williams T. Language-capable robots may inadvertently weaken human moral norms. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) pp. 401–410 IEEE; 2019.
98. Banks J. A perceived moral agency scale: development and validation of a metric for humans and social machines. Comput. Hum. Behav. 2019;90:363–371.
99. Scheutz M., Malle B., Briggs G. Towards morally sensitive action selection for autonomous social robots. In: 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) pp. 492–497 IEEE; 2015.
100. Chakraborti T, Kambhampati S. Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-AI collaboration. 2018. arXiv:1801.09854.