ARTIFICIAL INTELLIGENCE

In situ bidirectional human-robot value alignment

Luyao Yuan[1]*†, Xiaofeng Gao[2]†, Zilong Zheng[1,3]†, Mark Edmonds[1]*, Ying Nian Wu[2], Federico Rossano[4], Hongjing Lu[2,5]*, Yixin Zhu[2,3,6]*, Song-Chun Zhu[1,2,3,6]*

A prerequisite for social coordination is bidirectional communication between teammates, each playing two roles simultaneously: as receptive listeners and expressive speakers. For robots working with humans in complex situations with multiple goals that differ in importance, failure to fulfill the expectation of either role could undermine group performance due to misalignment of values between humans and robots. Specifically, a robot needs to serve as an effective listener to infer human users' intents from instructions and feedback and as an expressive speaker to explain its decision processes to users. Here, we investigate how to foster effective bidirectional human-robot communications in the context of value alignment—collaborative robots and users form an aligned understanding of the importance of possible task goals. We propose an explainable artificial intelligence (XAI) system in which a group of robots predicts users' values by taking in situ feedback into consideration while communicating their decision processes to users through explanations. To learn from human feedback, our XAI system integrates a cooperative communication model for inferring human values associated with multiple desirable goals. To be interpretable to humans, the system simulates human mental dynamics and predicts optimal explanations using graphical models. We conducted psychological experiments to examine the core components of the proposed computational framework. Our results show that real-time human-robot mutual understanding in complex cooperative tasks is achievable with a learning model based on bidirectional communication. We believe that this interaction framework can shed light on bidirectional value alignment in communicative XAI systems and, more broadly, in future human-machine teaming systems.

INTRODUCTION

At the dawn of artificial intelligence (AI), Wiener (*1*) identified the foundation of collaborative robots with the warning "if we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively ... we had better be quite sure that the purpose put into the machine is the purpose which we really desire." Since then, several efforts (*2*, *3*) have demonstrated that effective human-robot collaboration depends on a shared team mental model that includes values (*4*), goals (*5*), and current states of the task (*5*). To achieve a shared team mental model, humans use communication as an efficient tool to establish a common team understanding of task expectations, with team members adopting anticipatory information-sharing strategies to accomplish collaborative tasks (*6*, *7*). In most cases, the sharing process is bidirectional among collaborators, because each teammate needs to fulfill the roles of both speaker and listener (providing private task-relevant information to partners while also accurately comprehending teammates' messages). Successful communication in human-robot collaboration can be signaled by bidirectional value alignment, with robots accurately inferring human values, combined with effective explanations of the robot's behavior to humans. If these prerequisites are not met, the collaboration may encounter unforeseeable difficulties due to erroneous expectations of teammates (*8*). Thus, for robots to become beneficial collaborators in human society, they must be receptive listeners and expressive speakers when interacting with their human teammates.

From the listener's perspective, algorithms such as inverse reinforcement learning (IRL) (*9*) combine human interactive data with conventional machine learning methods to learn human values in specific tasks (*10*, *11*). Assuming (sub-)optimal behavior from human experts, IRL aims to recover the underlying reward function that guides human demonstration. However, acquiring human data in some application domains that arise in military and health care contexts can be expensive, if not impossible. Dependence on large datasets also prevents these methods from tackling in situ, real-time, and interactive human-robot collaboration scenarios. From the speaker's perspective, explainable artificial intelligence (XAI) was introduced to facilitate the alignment of mental models between humans and robots (*12*). However, existing XAI systems typically emphasize the generation of interpretable rationales to explain model decisions or predictions, either unfolding the model for a human user to probe and inspect (*12*–*16*) or reconciling the discrepancy between the human user's mental model and the robot's counterparts for a world model (*17*, *18*) and goals (*19*, *20*). Critically, human users' active interactions or inputs to the system only influence how explanations of robots' decisions are generated but rarely influence the model's decision-making process. This amounts to a unidirectional alignment of the mental model as static machine–dynamic human communication, where only the human user's comprehension of the robot or the task evolves given explanations about a fixed decision model in machines. In a nutshell, existing XAI systems primarily approach the human-robot communication problem from one of the two communication directions, but seldom from both. To accomplish bidirectional human-robot mental alignment, a more human-centric, dynamic machine–dynamic human communication is required. In such a paradigm, a robot, in addition

[1]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095, USA. [2]Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095, USA. [3]Beijing Institute for General Artificial Intelligence (BIGAI), Beijing 100080, China. [4]Department of Cognitive Science, University of California, San Diego, San Diego, CA 92093, USA. [5]Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095, USA. [6]Institute for Artificial Intelligence, Peking University, Beijing 100871, China.
*Corresponding author. Email: yuanluyao@ucla.edu (L.Y.); markedmonds@ucla.edu (M.E.); hongjing@ucla.edu (H.L.); yixin.zhu@pku.edu.cn (Y.Z.); sczhu@stat.ucla.edu (S.-C.Z.)
†These authors contributed equally to this work.

to revealing its decision-making process, would adopt the user's values and change its behavior in real time so that the robot and the human user would cooperatively achieve a set of common goals. To grasp the user's messages instantaneously, conventional data-driven machine learning approaches are replaced by communicative learning within a cooperative team. Explanations from the robot will be contextually adapted according to the human's current goals. Such a cooperation-oriented human-machine teaming would require the machine to have a certain level of Theory-of-Mind (ToM): A machine would actively infer the user's beliefs, desires, and goals (*21*, *22*). The system's design will not be limited to explaining its decision-making process but will also aim to understand human needs for cooperation, therefore forming a human-centric and human-compatible process (*23*). This mental alignment process, which can be viewed as one core computation for forming communal and personal common ground (*24*) that guarantees the coherence of human conversations, facilitates the success of human-machine collaboration.

Motivated to build an XAI system with the aforementioned capabilities of understanding the human user's beliefs, desires, and goals while being interpretable to the user, we introduce a sequential decision-making task that requires human-machine teaming to deal with complex constraints over problems intractable to the human's inferential capabilities. Specifically, we devise a human-machine teaming system instantiated as a collaborative game, in which the human user needs to work together with a group of robot scouts to accomplish some tasks and optimize the group gain. In this game, the user and robots communicate on a constrained channel. Only the robots directly interact with the physical world. The user does not directly access the physical world or directly control robot behavior. Only the user has access to the ground-truth value that encodes human's desirable end states, which determine how the task should be completed (for example, minimizing time and maximizing areas to explore), and the robots have to infer this value function through human-machine interactions. Such a setting constitutes a miniature task that realistically mimics real-world human-machine teaming. Many systems perform autonomously and interact directly with the hazardous environments under human users' supervision, but it is challenging (*25*) for desirable end states to be explicitly coded in autonomous agents beforehand or to change dynamically as events unfold. This setting also follows the classic multiagent system collaboration framework, where agents in the system can work in parallel but may rely on their partners' communication and feedback (*26*). To complete a game successfully, robots are expected to accomplish bidirectional alignment by both "listening" and "speaking" wisely. First, robots need to extract useful information from human feedback to infer the user's values and adjust their policies accordingly. Second, robots are required to effectively explain what they have done and plan to do based on their current value inference so that the user knows whether the team shares the human values. Figure 1 illustrates the bidirectional value alignment process in the game. Together, the proposed XAI system aims to address the following two questions. How can robots accurately estimate users' intentions during real-time interaction and feedback? How can robots explain themselves so that the user can understand their behavior and provide helpful feedback to aid their value alignment?

To learn human values and intentions, robots make proposals for task plans and ask for the user's feedback (acceptance or rejection of a proposal), from which the task goals can be inferred. In the collaborative game, knowing that robots are actively learning human values, the user tends to provide helpful pedagogical feedback to facilitate alignment (*27*). In particular, every message conveys two aspects of meanings, including literal meaning based on consistency between this message and the value and pragmatic meaning (*28*–*30*) based on deficiency of alternative feedback. Aware of the user's helpfulness, the robots adopt a human-centric amelioration of iterative teacher-aware learning (ITAL) (*31*) to learn the human value. ITAL performs maximum likelihood estimation (MLE) based on a two-part likelihood function: The first part models the probability of given feedback being aligned with the human value (literal meaning), and the second part captures the probability of receiving that feedback instead of other alternatives (pragmatic meaning). Leveraging both aspects of meanings, the proposed XAI system demonstrates value alignment in an in situ, few-round, instantaneous manner, enabling interactive human-machine communication in a cooperative teaming task with a large problem space. To synchronize the robots' mental status with the human user, our XAI system generates explanations that reveal robots' current estimation of human values and justify the proposed plan. In each step of interaction, to avoid overwhelming the user's cognitive workload with verbose explanations, the robots present customized explanations, such as omitting repetitive signals and emphasizing important updates. The robots model human users' mental dynamics as a Markov process and track the most relevant aspects of the robots' decision process using a sequential statistical graphical model. The explanation that includes all the relevant aspects and best addresses the user's concern at that step will be presented. After receiving explanations from robots and sending feedback to them, the user provides cues to the robots about how satisfying they found the latest proposals and explanations. Using this feedback, the robots constantly update the formats, attention, and contents of the explanations.

To evaluate the performance of our XAI system, we conducted human experiments to examine the success of bidirectional human-robot value alignment. We adopted three types of explanations and randomly assigned participants into one of the three groups. Three dependent measurements were used to assess the mental accordance, including the consistency between robot's inferred value and human's true value, human perception of how well robots infer and align with human's value, and human's cognitive trust (*32*) of the system. Our results show that the proposed XAI system can achieve bidirectional value alignment in an in situ, real-time manner for collaborative tasks; the robots can infer the human user's values and make their value estimation comprehensible to the user. We also found that some forms of explanations that benefit the way humans interact with the robots may not necessarily improve the human perception of how well robots infer users' values. These results provide converging evidence supporting the necessity for diverse explanations that promote both the performance quality of robots and their social intelligence (*33*). Because the goal of an AI collaborator is to reduce the human's cognitive burden and assist task completion, we believe that proactively inferring human values in real time and fostering human comprehension of the system pave the way for generic human-machine teaming.

RESULTS

Figure 1 illustrates the bidirectional value alignment procedure between the human user and robots during the game. The system's

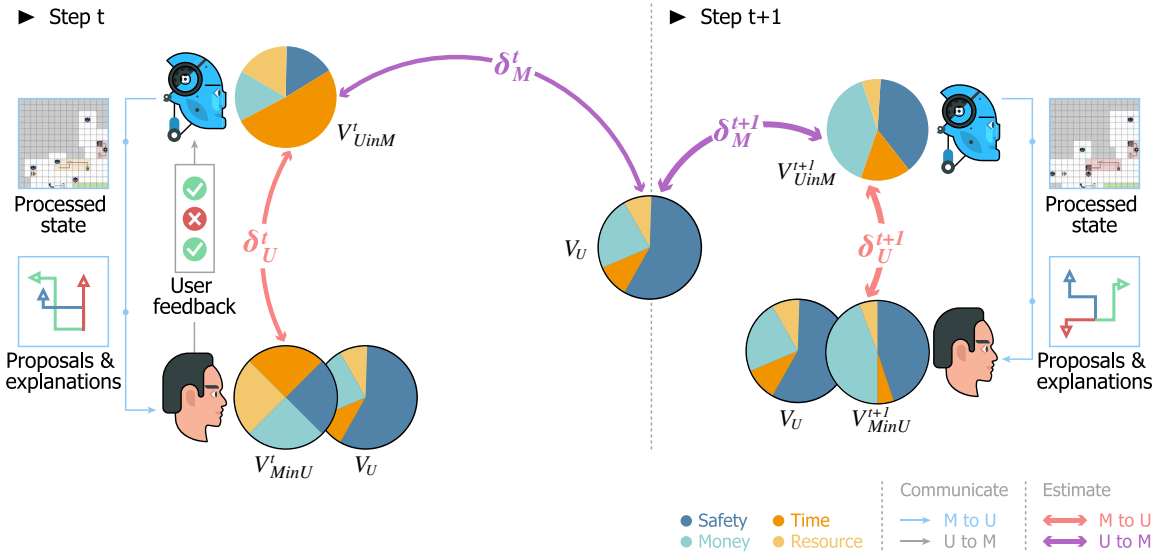

Fig. 1. Overview of bidirectional human-robot value alignment. Pie charts represent the values, which are the importance of different goals in a collaboration task, such as simultaneously considering safety, gaining money, saving time, and reserving resources. *t* in the superscript represents the time step. U and M in the subscript stand for "user" and "machine," respectively. V_U is the user's true value, V_{UinM} is the robot's estimation of the user's value, and V_{MinU} is the user's estimation of the robot's current value. δ denotes the distance between values in the task value space. In every round of interaction, the machine first receives signals from the physical environment and processes its observations to form an abstract state of the environment. Next, the machine presents the processed map together with movement proposals and explanations to human users, who will provide feedback to the system accepting/rejecting the proposals according to human values and current map state. Given the user's feedback, the machine then updates its estimation of human values and takes actions with respect to the new values. Cooperative human-robot communication with appropriate explanation aligns the team values in two directions by diminishing the distance between V_{UinM} and V_U, as well as V_{MinU} and V_{UinM}, resulting in final convergence to the true value V_U.

learning algorithm, built on top of ITAL (*31*), substitutes the conventional likelihood functions for regression or classification tasks by explicitly integrating the Boltzmann rationality human-decision model. The system incorporates both the literal and the pragmatic meaning of human feedback to infer the user's value. Meanwhile, the system explains its decision-making process to facilitate human perception of the machine. To test the system, we designed a psychological experiment to assess the performance of human-robot value alignment using different forms of explanations. In the coming sections, we first describe the human-robot collaboration game design, followed by an overview of the algorithms we used for bidirectional value alignment and explanation generation. Next, we introduce the human experiment design, report the empirical results of value alignments between humans and machines, and compare the effects of various types of explanations on value alignment.

Game design

Our collaborative game, the Scout Exploration Game, involves one human commander and three robot scouts. The game's objective is to find a safe path on an unknown map from the base (located in the bottom right corner of the map) to the destination (located in the upper left corner of the map). The map is represented as a partially observed 20 tile–by–20 tile board, with each tile potentially holding one of the various devices and remaining unobserved until a robot scout moves close enough to observe (reveal) the tile's contents. Every scout has the same probabilistic observation model, specified later in the "Observation model" section.

There is structural interdependence between human and scouts in the game (*34*). On the one hand, the user depends on the scouts to explore the dangerous area and defuse bombs. On the other hand, the scouts need the user to provide feedback to better understand the goal of the current mission. We define a set of goals for the robot scouts to pursue as they find the path to reach the destination, including saving time used to reach the destination, investigating suspicious circuits/bombs (loops of devices connected with wires) on the map, exploring tiles, and collecting resources (gold bricks on the map). The game's performance is measured by the accomplishment of these goals by the robot scouts and their relative importance (weights), defined as the human user's value function. Although all goals have intrinsic benefits, a trade-off among the goals has to be made according to the value function. For instance, if time is valued more in the value function than resources, the scouts should ignore some resources along the way to the destination for the sake of time. To emphasize the trade-off essence of value functions, we represent the importance of each factor with a percentage and the four percentages sum to 1. Before the interaction begins, a value function is assigned only to the human user as the mission for the game. Just like in the real world, various tasks can be specified with distinctive functions defined on a unified set of features (*35*). We coined seven value functions in this study to cover diverse types of tasks: Four of them have one dominant goal, two of them have two equally important goals, and the rest values everything but resources.

We mimic a realistic scenario, where human needs can be too diverse to code in the robot beforehand and value functions can be difficult to transfer between human and machine due to different mental representations. Without knowing the value function, to complete a task, the robot scouts (as a team) must quickly infer the commander's value. In each step, we let the robot team make three movement proposals, one for each scout, to the user, and the user can either accept or reject a proposal. To help the commander make

decisions, the robot team also explains the reason for every proposal. With the user's feedback, conditioned on the interaction history as well as the current map status, the robot team adjusts its estimation of the human value and takes actions accordingly. Specifically, if a plan is accepted, the proposer will follow that plan as much as possible (a plan may be interrupted by unexpected blocks in the partially observable map); otherwise, the robot will execute a new plan with the updated value estimation. We only allow the robot team to make proposals once in every round so that they must rely on their own autonomy to complete the task, instead of proposing until acceptance and effectively being teleoperated by the user. This concludes one round of interaction, and this process will be repeated. To avoid inefficient communication, the robot only makes proposals when necessary and acts according to the basis of the latest human value estimation. Figure 2 summarizes the human-machine interaction flow.

This game is complex through the lens of the combinatorial game theory. With an average planning step of 35 and a branching factor of 2^{12}, the estimated game tree complexity is 10^{126} for scouts to generate task plans. In comparison, chess has a game tree complexity of 10^{123}. On the player side, with an average round of feedback of 18 and a branching factor of 8, the estimated game tree complexity is 10^{16} for the player to provide feedback.

Bidirectional value alignment

Bidirectional value alignment, as one of the primary contributions, provides a more human-centric, dynamic machine–dynamic human communication framework for human-robot teaming. To estimate the human user's value during the communication process, we integrate two levels of ToM into our computation model. The level-1 ToM encodes the cooperative assumption. Namely, given a cooperative human user, the accepted proposals are more likely to align with the correct value function than the rejected ones. The level-2 ToM further accommodates users' pedagogy into the model. That is, the feedback that drives robots' value closer to the true value is more likely to be selected than other alternative feedback combinations. The pedagogical inclination requires an additional level of ToM because it demands recursive modeling of the user's model of the robots. Combining both levels of ToM, we formulate human behavior with distributions parameterized by the value and develop a learning algorithm with a closed-form parameter update function; see details in the "Human-robot value alignment" section.

To facilitate such a bidirectional alignment and gain human trust, we provide different forms of explanations along with proposals, which unveils the rationales behind scouts' proposals. Specifically, the explainer takes in current estimations of two levels of ToM as semantic input and fills it in a syntactic template. To provide concise explanations that are interpretable to humans and facilitate learning from humans, we devise a sequential generation process that selects templates by taking human's preference (reflected by the satisfaction score) over previously observed explanations into consideration. We call such preferences human's explanation utility; see details in the "Utility-aware explanation generation" section.

Human experiment

Experimental design

The human study examines whether our XAI system achieves real-time bidirectional value alignment between the human and the machine. In particular, we evaluate the efficacy of different forms of explanation of the robots' plans to human users. We conducted a psychological study with 135 participants. Participants were randomly assigned to one of three groups, including a proposal-only group, a brief-explanation group, and a full-explanation group. Each group has 45 participants. In the proposal-only group, the scouts only make proposals and give no explanations to the human. In the brief-explanation group, every proposal consists of one brief sentence explaining its positive outcomes. In the full-explanation group, a more detailed explanation accompanies every proposal, expounding the gains and costs of scouts' tentative actions and the dynamics of their values for the importance of different goals. Across all three groups, the robot scouts follow the same action policy and decision process for belief updating. The three groups differ only in terms of the forms of explanations provided to the human participants. Figure 3 compares the game interface that appeared in each group.

Our experimental setup consists of three phases: introduction, familiarization, and game playing. The first two phases prepare participants for the game. During the game, participants were asked to accept or reject scouts' proposals and assess satisfaction with the scouts' communication after every feedback. The feedback for proposals was given using buttons shown in Fig. 3B. The satisfaction assessments were provided via Likert-scale questions shown in Fig. 5A. In addition, we also asked participants to estimate the machine's internal states, such as the scouts' current value function and their qualitative trust of the XAI system.

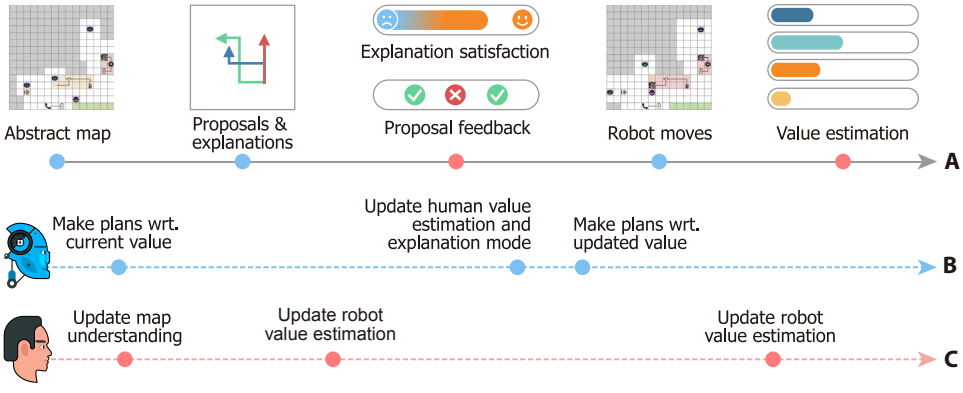

Fig. 2. Study design of the Scout Exploration Game. Timeline (**A**) denotes events happening in a single round of the game, starting from scouts receiving environment signals and ending with their next move. Proposals and explanations are presented differently to users depending on their experimental group (Fig. 3). The value estimation asks users to infer scouts' value at the current time. Answers to these questions are not used by the scouts during the game, but only for inspecting users' mental model after the game is complete. Figure 5D shows the detailed UI of these questions. Timelines (**B**) and (**C**) depict mental dynamics of the robots and the user, respectively.

Human study analysis

Figure 4 illustrates the bidirectional human-robot value alignment results for all three groups. We compute the Kendall rank correlation coefficient, commonly referred to as Kendall's τ coefficient, to assess the value alignment between scouts and humans. Perfectly

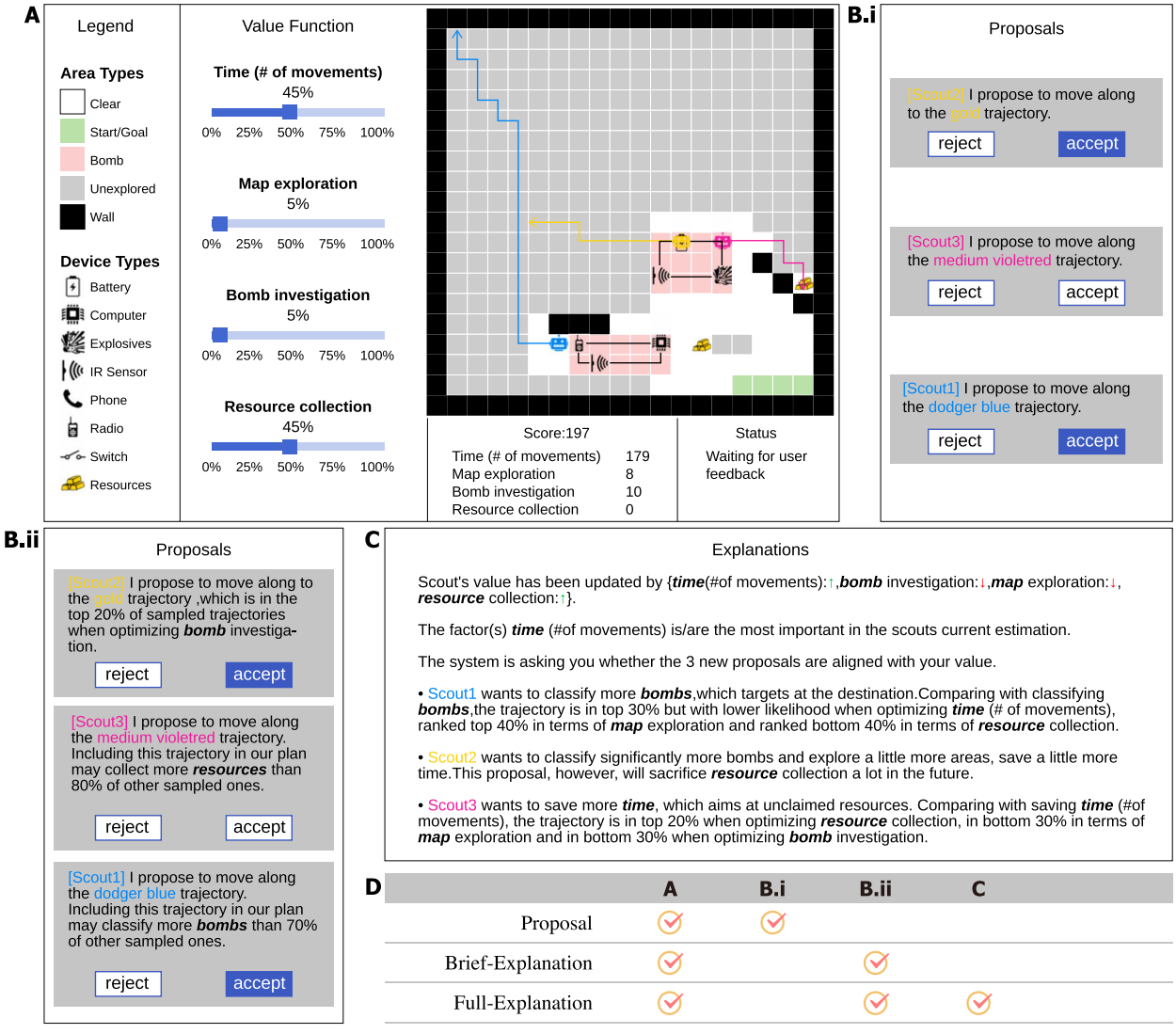

Fig. 3. UI for the Scout Exploration Game. (**A**) From left to right: The Legend panel explains the meaning of all icons used in the game; the Value Function panel shows the true values indicating the relative importance of various goals, which is unknown to the scouts; and the map panel shows the current status of the map in the game, including the grid map, the current scores for achieving individual goals, and the current status of the robot system. (**B**) The Proposals panel shows the robot scouts' current proposals; human users can accept or reject proposals of individual scouts. In the proposal-only group, participants only see a descriptive sentence for each proposal (B.i), whereas, in the brief-explanation and full-explanation groups, participants are presented with a brief explanation about the proposal's purpose (B.ii). (**C**) The Explanations panel shows detailed explanations provided by the scouts, only displayed to the full-explanation group. (**D**) The table summarizes key components of the game display included in each group.

agreed/disagreed rankings have $\tau = \pm 1$, and independent rankings expect $\tau \approx 0$. To demonstrate dynamic changes in bidirectional value alignment between scouts and humans, we recorded the scouts' estimation of the user's value and measured the user's estimation of the scout's value as the game proceeds.

Figure 4A shows the alignment between robots' estimated values and the true values known to human users. First, all groups show higher value alignment at the end of the game compared with the beginning of the game [paired t test, $t_{\text{Prop}}(44) = 2.850$, $P_{\text{Prop}} = 0.007$, $t_{\text{Brief}}(44) = 10.148$, $P_{\text{Brief}} < 0.001$, $t_{\text{Full}}(44) = 11.452$, $P_{\text{Full}} < 0.001$]. Scouts that interacted with the brief-explanation and full-explanation groups show stronger value alignment, revealed by higher correlations between scouts' estimated values and true values, than alignment in the proposal-only group ($\tau = 0.2$, 0.4, and 0.5 for the proposal-only, brief-explanation, and full-explanation groups, respectively). The group differences emerge in early stages of the game (25% of the game progress) and are maintained to the end of the game, confirmed by analysis of variance (ANOVA) at a range of progress points [from game progress 25, 50, 75, and 100%, $F_{(2,132)} = 19.086$, 14.202, 11.961, and 11.622; $P < 0.001$, $P < 0.001$, $P < 0.001$, and $P < 0.001$]. Better value alignment in the two groups involving explanation than the baseline proposal-only group without explanation provides strong evidence that explanations about robot decision processes to human users enhance bidirectional communications between humans and machines. The enhanced communication, in turn, helps machines gain accurate estimates of

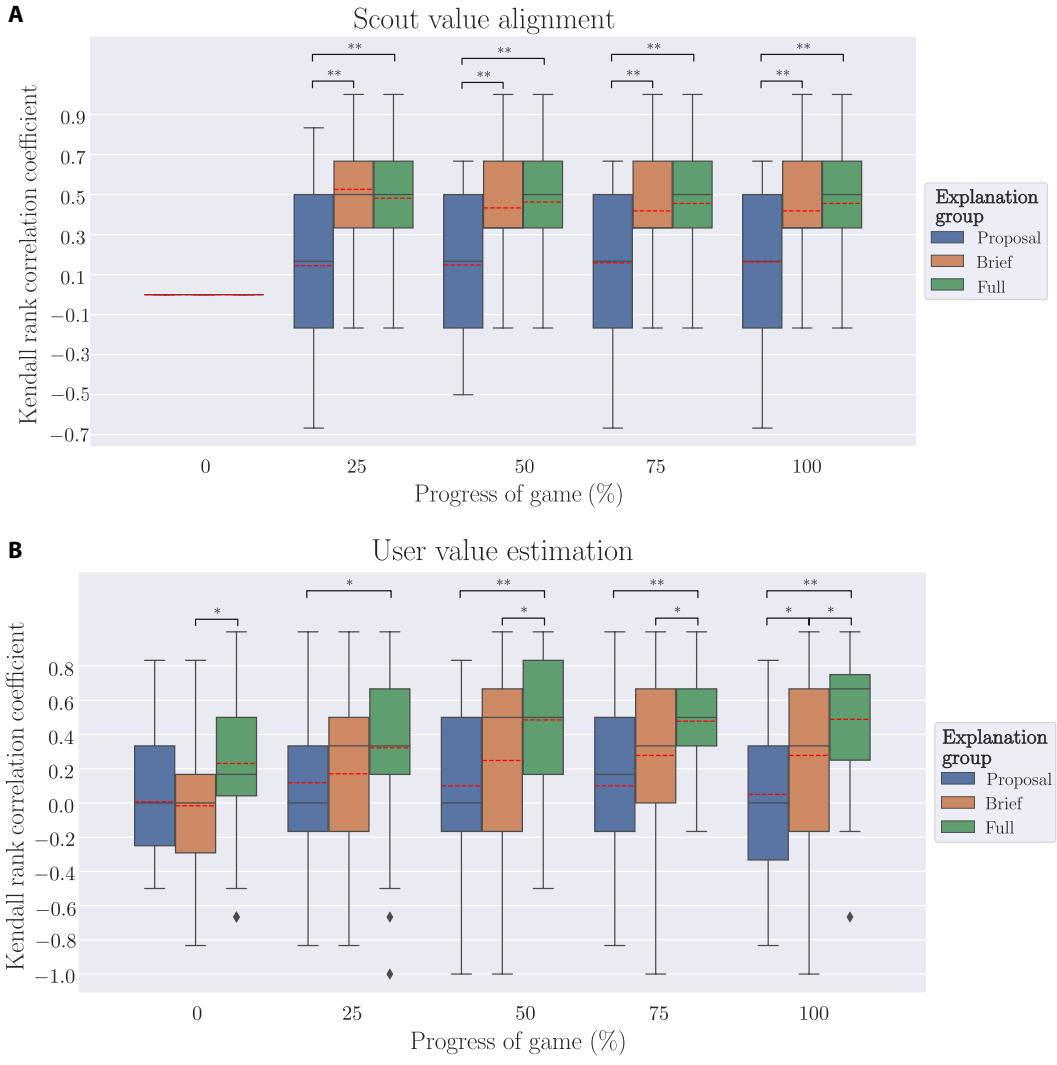

Fig. 4. Box plots showing results of value estimation for scouts and humans in three groups. The legends: Proposal, Brief, and Full refer to the proposal-only group, the brief-explanation group, and the full-explanation group, respectively. Horizontal axis indicates the progress of the game for human participants; vertical axis indicates Kendall's rank correlation coefficient between estimated values by scouts and humans; higher correlation indicates better value alignment. (**A**) Correlation between scouts' value estimate and the true values that are known to human users as a function of game progress. It represents the scout's accuracy in estimating human values. Before the game starts, the scouts' value estimate is initialized as uniform across all goals. (**B**) Correlation between the human estimate of the scouts' values and scouts' estimate of the true values as a function of game progress. It represents humans' accuracy in estimating scouts' values. * indicates significant group differences in paired *t* test with *P* value smaller than 5% and ** indicates that *P* value is smaller than 1%. The solid lines and red dashed lines in the bars respectively indicate the median and mean.

human values, thereby fostering human-machine teaming. There is no difference between the brief-explanation group and full-explanation group, implying that the detail of the explanations may not critically influence humans' feedback in terms of accepting or rejecting robots' proposals, as long as these explanations provide sufficient contexts to justify the robots' intents. Figure 4B depicts how well the human users estimate the scouts' values over the progress of the game. It represents the accuracy with which humans assess the scouts' values. Figure 5D shows the interface we used to collect human estimates in the experiment. An ANOVA test revealed a significant main effect of groups in the later stage of the game playing at game progress 50, 75, and 100% [$F_{(2,132)} = 7.632$, $F_{(2,132)} = 8.339$, and $F_{(2,128)} = 10.542$; $P = 0.001$, $P < 0.001$, and $P < 0.001$, respectively]. The brief-explanation and full-explanation groups show significant enhancement of alignment between human estimates of scouts' values and scouts' values used in determining their decision and proposals at the end of the game compared with the beginning of the game [paired *t* test, $t_{\text{Brief}}(44) = 3.272$, $P_{\text{Brief}} = 0.002$, $t_{\text{Full}}(39) = 2.810$, $P_{\text{Full}} = 0.007$], whereas the proposal-only group does not show any improvement [paired *t* test, $t_{\text{Prop}}(38) = 0.286$, $P_{\text{Prop}} = 0.776$]. These results suggest that human users have difficulty understanding robots' intentions by only observing their situational behaviors, highlighting the central role of explanation in revealing the robots' intentions to its human user. Critically, humans show stronger alignment in estimating scout's values in the full-explanation group than in the other two groups in the second half of the game [independent sample *t* test, from game progress 50, 75, and 100%, $t(88) = 4.291$, $t(88) = 4.511$, and $t(84) = 5.088$, $P < 0.001$, $P < 0.001$, and $P < 0.001$ against

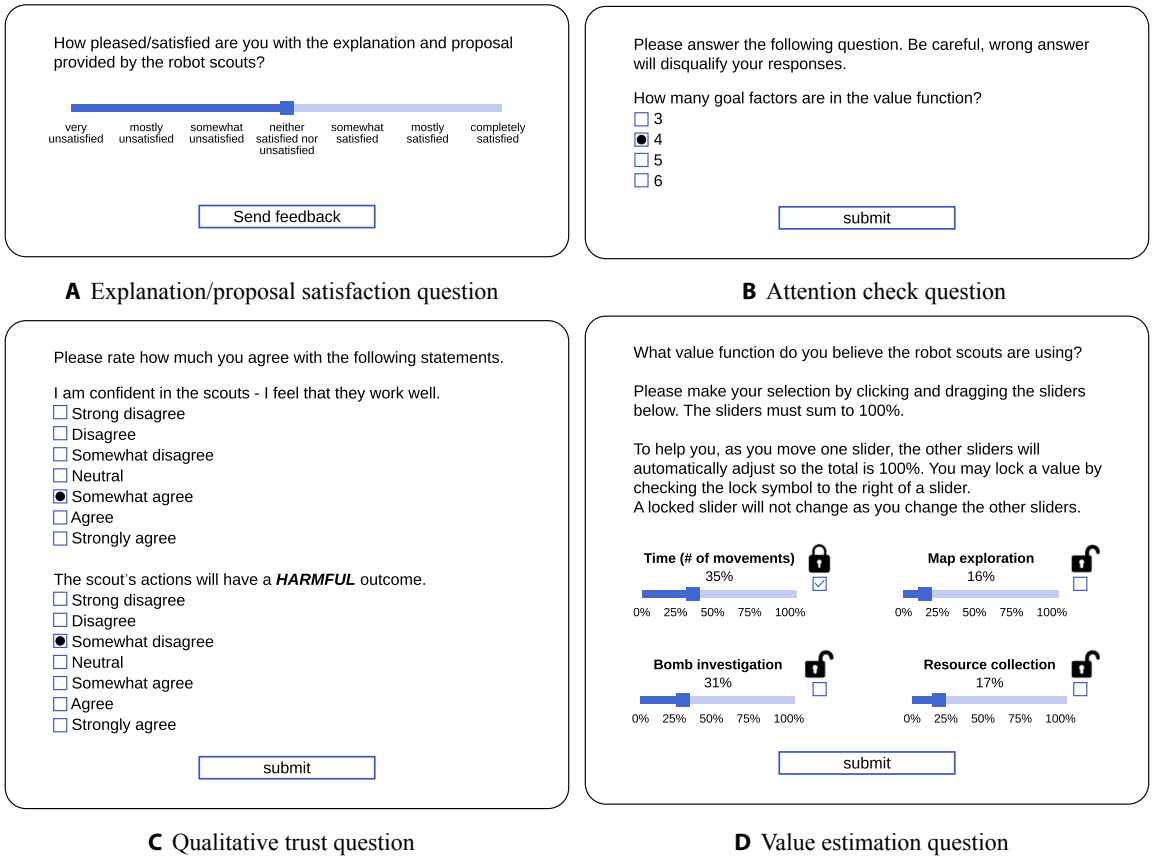

Fig. 5. Examples of questions participants received during the game. (**A**) Explanation/proposal satisfaction question. Participants are asked to provide a satisfaction score for the explainer in every round when they receive scout's proposals and explanations. This satisfaction score is used to update models for generating future explanations. (**B**) Attention check question. These questions are shown after trust questions; participants receive one of the four questions about the game logic and UI. Participants who failed the attention check are later removed from data analysis. (**C**) Qualitative trust question. We ask the participants "how confident you are in the scouts?" and "how much do you think the scout's actions will have a harmful outcome?" (**D**) Value estimation question. Participants predict the robot scouts' belief about the true human value by sliding the bars to set a relative importance of each goal; this is a question about level-2 ToM. Our interface ensures that the total value of all goals sums to 100%; if the participant moves one slider, the others will automatically change proportionally with respect to their original values such that all values still sum to 100%. Meanwhile, participants can lock a particular slider by checking the lock symbol to the right of the slider.

proposal-only; $t(88) = 2.387$, $t(88) = 2.219$, and $t(86) = 2.196$, $P = 0.019$, 0.030, and 0.031 against brief-explanation]. In comparison, the brief-explanation group only yields a more consistent human estimate of scouts' values than the proposal-only group at the end of the game [independent sample t test, $t(86) = 2.274$, $P = 0.026$]. Together, these results indicate that both forms of explanation facilitate human estimates of robots' value estimates based on observation of robots' behavior and interactions with the robots. However, the full explanation, which provides details about both the advantages and the disadvantages of a proposal, is more helpful to human judgments about the robots' estimates than a brief explanation showing only the major benefit of a proposal.

Both results from robots' estimate of human values (Fig. 4A) and from human estimate of robots' values (Fig. 4B) show that groups with brief and full explanations can maintain a stable trend of value alignments across the entire human-robot teaming process. However, the emergence of value alignment differs across different groups and varies for different alignment metrics over the course of the game. Robots' value alignment metrics measured by the Kendall's τ coefficient converges at 25% of the game. Alignment of human estimates and scouts' values converges at 50% progress for the full-explanation group and at 75% game progress for the brief-explanation group. These results demonstrate our system's capability to maintain the established team mental model during continuous human-robot teaming, with full explanations enabling faster convergence of users' estimates of robots' mental status (estimates of values). The convergence of both alignment metrics shows that our value alignment algorithm enables the robot scouts to learn human values in an in situ, real-time, and interactive manner. It also shows that explanations generated by the robots enable users to better perceive the machine's values. These results demonstrate a bidirectional human-robot alignment. Moreover, our result pins down the contributions of explanation formats in different facets of human-robot communication. We found that brief and full explanations lead to similar effects in improving the way humans provide feedback to the machine via acceptance or rejection of robots' proposals. However, the full-explanation group shows a significantly greater benefit for human accuracy in estimating robots' values.

DISCUSSION

Our proposed XAI system successfully demonstrates the feasibility of a bidirectional human-robot value alignment framework. From the listener's perspective, robots in all three explanation groups can quickly align to the user's value by correctly ranking at least 60% of goals' importance as early as the 25% progress of the game. From the speaker's perspective, by providing proper explanations, robots can reveal their intentions to the user and facilitate better human perception of the machine's values, with convergence occurring at 50% (full-explanation) and 75% (brief-explanation) of the game. Together, both perspectives provide convincing evidence of a bidirectional process of value alignment. On the one hand, by receiving cooperative human feedback, robots gradually update their value function to align with the human values. On the other hand, by continuously interacting with the robots, the human user gradually forms a coherent perception of the system's capability and intentions. Although the system's values have not converged in the first half of the game, the user's perception of the robots' estimate can still improve. Eventually, when the robots' values become stable, the user's estimation of the robots also becomes stable. The pairing of convergence from robots' estimate of the user's values to user's true values and from user's estimate of the robots' values to robots' current values forms a bidirectional value alignment anchored by the user's true value.

Despite showing similar converging trends of value alignments, the three explanation groups differ in the precision of their alignments. In both directions of human-robot estimation, including scouts estimating human values and the human's understanding of the scouts' current value estimation, the Kendall's τ coefficients of the proposal-only group are significantly lower than the coefficients of the other two groups. These gaps suggest that human-machine interactions alone are not sufficient to enhance the human perception of the machine, nor are they sufficient to evoke better human feedback/guidance to the robots. Results from the computational ecology models show that a multiagent system can converge to an equilibrium point only when the information delay and uncertainty between agents are fairly small. On the contrary, our modeling framework can handle relatively large amount of uncertainty. In our case, the scouts' explanations play an important role in reducing information uncertainty and system convergence: Explanations help the human understand the machine's current value estimation and generate a better response, which in turn enables the machine to estimate the value more accurately. The extent of human-robot mutual understanding depends upon how an AI system explains itself to the user. In the absence of informative explanations, certain misconceptions cannot be eliminated even with continuous interaction between humans and robots, leading to a slower value alignment.

Compared with the brief-explanation group, the full-explanation group demonstrates significant enhancement in the human estimate of robots' values but does not show a strong advantage in terms of scouts' value alignment. These results indicate that human users in the two explanation groups provide feedback to the scouts in similar ways, although the full-explanation group acquires a more accurate understanding of the system. One possible cause of this dissociation is that human users exhaust their cognitive resources when processing other complexities of the game, such as comprehending messages from the three scouts or analyzing information on the map. Thus, additional details in the full explanation cannot be accommodated to offer more rationale feedback. An alternative possible reason involves the design of the game, in particular, the granularity of the feedback. Because there are only eight possible feedback combinations (two for each scout's proposal) in every round, it is possible to identify the best response to the scouts with only a limited understanding of the system status and the proposals. The extra information provided in the full explanation, although beneficial to the user's perception of the machine, maybe useless for scouts' value learning.

We also measured users' qualitative trust in the system as the game proceeds. However, we did not find any significant differences across the three explanation groups. This result suggests that human trust toward machines depends on many facets of the machines (*33*, *36*). Both social intelligence and performance quality of robots are indispensable to fostering trust (*33*). Better bidirectional value alignment can improve team performance but may not be sufficient to enhance the human perception of scouts' social intelligence through short-term human-robot collaborations. In addition, in the current game, scouts are not likely to make catastrophic mistakes throughout the task and are guaranteed to reach the upper left corner of the map successfully. Because the robots can always accomplish the task in the end, users may tend to trust the robots so that explanations have less effect on trust formation.

To summarize, we present a bidirectional human-robot value alignment framework and use an XAI system to verify its feasibility. The proposed XAI system demonstrates that, with ToM integrated into the machine's learning module and appropriate explanations provided to the user, humans and robots are able to achieve alignment of mental models through an in situ, real-time, and interactive manner. The coherent computational framework reported in our study provides promising results to address the question "what constitutes a good human-robot team," by contributing to the formation of a shared mental model between a human and a machine. Particularly, our work focuses on the task-specific aspects of the mental model, namely, the value and intentions. In more intricate scenarios, mental alignments can further entail other aspects going beyond the context of a single task—for example, capabilities of every team member, prerequisites and outcomes of actions [also referred to as the world transition model in reinforcement learning (RL)], or individual duties and roles. These components in the mental model are useful across various task contexts. In human language, using such a mental alignment process is often referred to as personal common ground and can be established via episodic evidence (*24*), which is the actions or events the speakers are part of together. In our setting, the episodic evidence could be acquired from human-robot collaboration in multiple games, possibly with different value functions and maps. As such, the universal human model described in Eqs. 2 and 6 can be replaced by a customized model parameterized for individual person's characteristics.

In this work, we focus on the alignment of value functions, which captures the relative importance of a wide range of goals. Aligning the values can greatly help the human and machine establish common ground for task-oriented collaborations. The mental models of human-robot teaming in this work are primitive compared with the ones that emerged in human-human collaborations, wherein common ground formed and maintained from rich shared experience (*37*, *38*). Specifically, we assume limited communication bandwidth, a realistic setting that occurred in real-world applications. We consider our work as a first step toward a more general mental model alignment setting in human-machine collaboration. In future work, we plan to explore factors that can further enhance human users' trust (for example, enabling counterfactual queries to the robots),

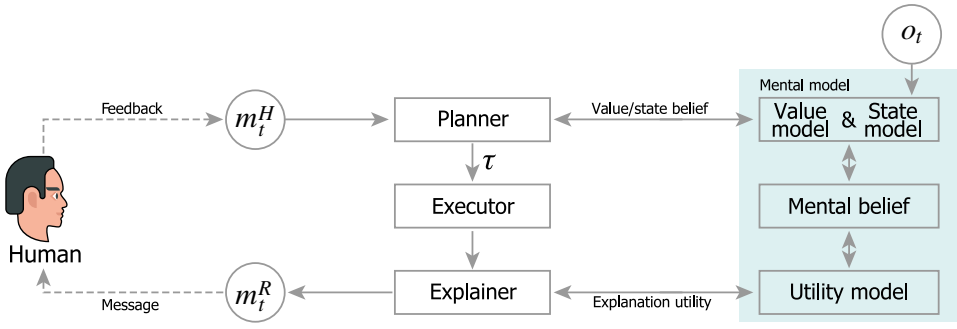

Fig. 6. Algorithmic flow of the computational model. Given game observations and human feedback to previous proposals, the robots update their mental state and make new task plans. On the basis of the plans and current beliefs, new proposals and explanations are generated and sent to the human for feedback in the next round.

validate the effects of alignment on task performance, and apply our system to tasks involving diverse communication modalities, more complicated environments, and richer mental states.

MATERIALS AND METHODS

Game setup

We implemented this game using HaxeFlixel, a two-dimensional game engine for JavaScript-based games, such that participants can access the game on web browsers; this setup was necessary in the situation of coronavirus disease 2019 (COVID-19). Our between-participant design is divided by the explanation format provided to the participants: the proposal-only group, the brief-explanation group, and the full-explanation group. The proposal-only group only shows the proposed trajectory on the map and a basic descriptive text about the proposal, such as "Scout 1 proposes to move along the blue trajectory." For the brief-explanation and full-explanation groups, brief explanations accompany the proposals to clarify the motivation of the robot scouts; for example, "Scout 1 proposes to move along the blue trajectory, which is in the top 1% of sampled trajectories when saving time." The full explanation includes a more detailed full explanation besides the brief one in the proposal panel; for example, "Scout 1 wants to save more time at the cost of map exploration and resource collection." More details about explanations can be found in the "Utility-aware explanation generation" section. The full user interface (UI) of the game is displayed in Fig. 3 with the actual explanations used in the game.

After giving feedback to the scouts' proposals, participants were asked a few questions before the next round of explanation and proposing. These questions were, by the order of showing up, satisfaction about the latest proposal and explanation, value estimation, qualitative trust, and attention check. Only answers to the satisfaction questions were used by the system's explainer for explanation utility tracking; all other questions were used only for post-game analysis. Figure 5 includes some example questions queried during the game. To avoid overwhelming users, value estimation questions were asked every two proposals, and qualitative trust and attention check questions were asked every five proposals.

Computational model details

Before diving into the technical details of how the proposed robot scouts act, align value, and interact with the human user in a bidirectional communicative learning framework, we first provide an overview of the game flow and the notations of the computational model. We use R and H to denote the robot scouts and the human user, respectively. θ encodes the parameters of the value function, s is the physical state, υ is the utility of explanations, and $b(\cdot)$ is the belief over latent variables. $x^R = (b(s), b(\theta), b(\upsilon))$, the mental state (3, 5) of the robots (the robot team shares one mental state), depicts their current beliefs of all the unknown task-relevant variables. m is the message used for human-machine communication. In every round of the game, the robot scouts receive observations from the environment and make a task plan based on their current mental state. Next, they send messages (proposals and/or explanations) to the human user for feedback; this user feedback is used for robots' final movement plans in this round. Algorithm S1 sketches the high-level game flow, and Fig. 6 shows the computation pipeline for one round of human-machine teaming.

One comparable but different setting with our human-machine teaming framework is IRL (*39*). Nevertheless, IRL aims to recover an underlying reward function given prerecorded expert demonstrations in an offline passive learning setting. In contrast, the robot scouts in our setting are designed to learn interactively from scarce supervisions given by the human user. Crucially, our design requires the robots to actively infer the human user's value in real time and in situ as the task proceeds. Furthermore, to consummate a collaboration, the robot scouts not only must quickly comprehend the human user's intent but also elucidate themselves to ensure smooth communication with the human user throughout the entire game. In brief, the robots are tasked to perform value alignment by inferring the human user's mental model, actively making proposals, and evaluating the human user's feedback, which requires complex and recursive mind modeling of the human user.

In the coming sections, we will introduce how the robots select actions, make proposals, update belief of human user's value function, and generate communication messages. In the "Observation model and belief of states" section of Supplementary Methods, we describe how the robots process observations and update their belief of the states.

Action selection

Suppose the robot scouts already know about the human user's value function. The game simplifies to a partially observable Markov decision process (POMDP) setting, solvable by planning-based methods (*40*). Let τ_i denote the plan proposed by the ith scout and $\tau = \{\tau_1, ..., \tau_K\}$ as the complete plan of the scout group, where K is the number of scouts in the group. When constructing a plan, the scouts use the following policy:

$$\begin{aligned}
&\arg\max_{\tau\in\mathcal{T}} E_{s\sim b(s),\theta\sim b(\theta)}\left[\theta^T f(\tau,s)\right] \\
&\quad = \arg\max_{\tau\in\mathcal{T}} E_{s\sim b(s)}\left[f(\tau,s)\right]^T E_{\theta\sim b(\theta)}\left[\theta\right] \\
&\quad \approx \arg\max_{\tau\in\mathcal{T}} \bar{\theta}^T\left(\frac{1}{N_S}\sum_{n=1}^{N_S} f(\tau,s_n)\right) = \arg\max_{\tau\in\mathcal{T}} \bar{\theta}^T\overline{\left(f(\tau)\right)}
\end{aligned} \quad (1)$$

where $f(\tau, s)$ is the status of the four goals in the terminal game state given that current state is s and the scouts follow the plan τ; we call it the features of (τ, s). The first equality holds because s and θ are independent of each other in our setting. Given the dynamics of the game, f can be forward simulated in our planner such that the expectation of $f(\tau, s)$ can be approximated using Monte Carlo methods with N_S state samples, giving us $\overline{f(\tau)}$, the feature of τ. Instead of computing the full distribution, the agent only needs to keep track of the mean of the belief over human user's value function because we are using a linear model to calculate the gain of the game; we use $\overline{\theta}$ to denote the mean of $b(\theta)$. Because the space of all possible plans is too large $[(20 \times 20)^K]$ to be calculated exactly, we use heuristics to approximate the space of all possible plans by constructing another space T and select the optimal plan from it; see the "Scouts plan space construction" section in Supplementary Methods for details about constructing T. After a plan τ is determined, the joint action of all robot scouts is the first move of the generated plans $a^R = (\tau_1[0], \ldots, \tau_K[0])$.

Proposal selection

To improve user experience during the interaction and foster human trust, the robots ought to make good proposals at the proper time to collect users' informative feedback. In active learning (*41*), the query usually maximizes the expected information gain. However, such criteria of asking questions to an oracle cannot be applied to human-robot interaction (HRI). Critically, besides acquiring information from the human, robots' questions also ought to reveal their mental status to the user and gain their trust. A clear dilemma always exists: The proposal with the most expected information gain is usually the most uncertain one as well, querying of which easily leaves an unreliable impression of the system to the user and impairs the human perception of the machine's value. To tackle this issue, we designed our communicative learning framework such that the scouts will propose the optimal plan given their current estimation of the user's value. Such proposals can reveal the robot team's current mental status to the user for better human perception of the scouts, hence receiving more helpful supervision. For instance, if all three proposals ignore suspicious devices, but bomb exploration is an important factor, the user will be aware of the discrepancy between the scout's value and the intended value and adjust it with feedback. As a result, plans used for proposals are calculated in the same way as plans for action selection described in the last paragraph, only with $b(\theta)$ from the previous time step.

Human-robot value alignment

Level-1 ToM

The robot scouts need to estimate the human user's value from their interactions. In the collaborative game, the more a proposal facilitates goals with high values, the more it is likely to be accepted. Here, we refer to the ability to infer humans' value from their actions as level-1 ToM. Bearing level-1 ToM, the scouts can interpret the user's feedback and update the value estimation given the current map status. For example, if a trajectory toward a partially explored circuit is accepted, the scouts are likely to increase the value to bomb investigation and lower the other goals. We integrated level-1 ToM into our computation model and developed a learning algorithm with a closed-form parameter update function.

Belief update with level-1 ToM

Let $m^H(fb)$ denote the human user's feedback, which is a binary code, with the ith bit indicating the acceptance or rejection of the proposal from the ith scout. Assuming that the human user considers each proposal separately and follows a Bernoulli acceptance distribution (*42*), the likelihood function of the human user's feedback is

$$p(m^H(fb)|\,\tau;\overline{\theta}) = \prod_{i=1}^{K} p(m^H(fb)_i|\,\tau_i;\overline{\theta}) = \prod_{i=1}^{K} \frac{\exp(\beta_1 \overline{\theta}^T \overline{f(\tau_i)})^{m^H(fb)_i} \exp(\beta_1 \overline{\theta}^T \overline{f_{\neg}(\tau_i)})^{(1-m^H(fb)_i)}}{\exp(\beta_1 \overline{\theta}^T \overline{f(\tau_i)}) + \exp(\beta_1 \overline{\theta}^T \overline{f_{\neg}(\tau_i)})} \quad (2)$$

where

$$\overline{f(\tau_i)} = \sum_{\tau \in T:\tau_i \in \tau} \overline{f(\tau)}, \text{and } \overline{f(\neg\tau_i)} = \sum_{\tau \in T:\tau_i \notin \tau} \overline{f(\tau)} \quad (3)$$

That is, a proposal is more likely to be accepted if including it in the scouts' plan is more beneficial than excluding it when $\overline{\theta}$ is the value parameter. Given this likelihood function, we use MLE to learn $\overline{\theta}$ by maximizing $\log p(m^H(fb)|\,\tau;\overline{\theta})$ with respect to $\overline{\theta}$:

$$\overline{\theta} = \overline{\theta} + \eta \frac{\partial \log p(m^H(fb)|\,\tau;\overline{\theta})}{\partial \overline{\theta}} \quad (4)$$

where η is the learning rate and

$$\frac{\partial \log p(m^H(fb)|\,\tau;\overline{\theta})}{\partial \overline{\theta}} = \beta_1 \sum_{i=1}^{K} \left[m^H(fb)_i \overline{f(\tau_i)} + (1 - m^H(fb)_i) \overline{f_{\neg}(\tau_i)} \right] - \mathrm{E}_{m \sim p(m^H(fb)_i|\,\tau_i;\overline{\theta})} \left[m \overline{f(\tau_i)} + (1-m) \overline{f(\neg\tau_i)} \right] \quad (5)$$

where acceptance/rejection selects the feature of including/excluding τ_i and the expectation is taken with respect to the feedback distribution given current $\overline{\theta}$. The expectation computes the average feature if the plan, τ, is randomly accepted/rejected according to current $\overline{\theta}$. The difference between the user's designated feature and the expected feature forms the gradient. Because $\overline{\theta} > 0$ and $\|\overline{\theta}\|_1 = 1$, we perform MLE with the projected stochastic gradient ascent algorithm.

Level-2 ToM

Intuitive but limited, the comprehension of feedback endowed by level-1 ToM is constrained to its plain content (the literal meaning of the feedback). In human communication, messages often convey both literal and pragmatic meanings (*43*). In other words, one can acquire not only explicit information from what others said but also implicit information from what others did not say. A typical concretization is the Gricean Maxims of quantity (*28*) or the scalar implicature: When people say "I like drinking warm coffee," although the lexical meaning of "warm" is semantically close to "hot," they mean "not hot"; otherwise, people would have said "hot" directly (*44*, *45*). Similarly, the human user's selection of a certain combination of feedback but not other combinations can also help robot value alignment. To comprehend this process, it requires the robots to mentally simulate and plan based on human users' pedagogical tendency and belief about the robots' current plan. We refer to such a recursive inference ability as level-2 ToM.

Belief update with level-2 ToM

To enable level-2 ToM, robots need to conduct a recursive mental simulation in a counterfactual fashion and consider the advantage of the received feedback over others not being sent. Intuitively, suppose the user knows how the robots with level-1 ToM update the value given feedback; the more the feedback leads to changes toward the ground-truth value, the more it is likely to be selected. Computationally, the

level-2 robots first simulate a level-1 value update given all possible feedback. Next, the robots find a ground-truth value such that the update brought by the received feedback is better than the other alternative feedback. Mathematically, we formulate the human user providing feedback based on its anticipatory improvement following

$$q(m^H(fb)|\overline{\theta},\tau;\theta^*)=\frac{\exp\left(-\beta_2\left\|\overline{\theta}+\eta\frac{\partial\log p(m^H(fb)|\tau;\overline{\theta})}{\partial\overline{\theta}}-\theta^*\right\|^2\right)}{\sum_{m^{\widehat{H}(fb)}\in FB}\exp\left(-\beta_2\left\|\overline{\theta}+\eta\frac{\partial\log p(m^{\widehat{H}(fb)}|\tau;\overline{\theta})}{\partial\overline{\theta}}-\theta^*\right\|^2\right)} \quad (6)$$

where $\beta_2 \geq 0$ controls the extremeness of the Boltzmann rationality, η is the learning rate, and θ^* is the set of ground-truth parameters of the value function possessed by the human user. The intuition of this equation is as follows: The feedback from the human user is sampled from a soft-min distribution of the distance between the updated parameters given the feedback and the ground-truth parameters. The smaller the distance is, the larger the improvement brought by that feedback, and the larger the improvement is, the more likely the feedback is provided. Further analysis of the above distance can be found in the study of Liu *et al.* (*46*). Integrating this feedback function into our value learning algorithm, we can derive a new parameter update function:

$$\overline{\theta} = \overline{\theta}+\eta g(m^H(fb))+2\beta_2\eta^2\left(g(m^H(fb))-\mathrm{E}_{m(fb)\sim q(m(fb)|\overline{\theta},\tau;\theta^*)}[g(m(fb))]\right) \quad (7)$$

where

$$g(m(fb))=\frac{\partial\log p(m(fb)|\tau;\overline{\theta})}{\partial\overline{\theta}} \quad (8)$$

The first two terms in Eq. 7 are the same as the level-1 belief update, whereas the third term grasps the message's context by comparing the selected message against the also-runs and leverages the advantage to further update the belief. Notice that θ^* is unknown to the agent, so q in the expectation does not have an exact solution. Thus, we use $\overline{\theta}+\eta g(m^H(fb))$ as an approximation of θ^*. That is, we calculate level-1 ToM update on the parameters of the value function and take an additional gradient ascent step for level-2 ToM update. In this work, we always initialize scouts' value as uniform across all goals: $\overline{\theta}^0 = [0.25, 0.25, 0.25, 0.25]$.

The difference between robots with level-1 ToM and level-2 ToM is the likelihood function they used to model the user. A level-1 robot assumes that the user provides feedback only by thinking about how good the proposals are, whereas a level-2 robot is also aware of the pedagogical perspective of the human user in the collaborative game and accommodating the information of both the literal and the pragmatic meaning of user feedback.

Theoretically, the recursive reasoning between robots and the human user can continue infinitely with unlimited resources or up to a fixed point of convergence (*47*). In this work, we only model the human user as knowing the value update mechanism of scouts with level-1 ToM, a manageable extent of reasoning for human cognitive capability (*48*), which is also adopted by recent literature (*49*).

The effectiveness of this computational model in the Scout Exploration Game has been verified by the empirical results in the previous section. For other settings, in which task performance has a linear relationship with the value, as depicted by Eq. 1, the same model can be applied with minor modifications. For settings involving nonlinear value functions, the inner product in Eq. 1 is to be replaced, as well as the gradient function in Eq. 5. Still, the core computations in the algorithm, namely, the MLE learning of the value function and the level-1/2 ToM integration, remain the same.

Utility-aware explanation generation

We generate explanations to aid the human user in collaborating with the robots by accepting/rejecting specific proposals. Given trajectories produced by the planner, the explainer aims to generate human-like explanations that not only provide sufficient semantic information but also match the human user's syntactic preferences, namely, the explanation utility. Specifically, an explanation is defined by its semantic inputs and a set of syntactic rules. The former is produced by the planner, providing explanations regarding what. This includes the current observation, physical state, and belief over the value function. The latter is to provide explanations regarding how, which corresponds to user's explanation utility.

To quantitatively estimate the utility values, after each round, we used a Likert-scale questionnaire on explanation/proposal satisfaction (Fig. 5A). Answers to these questions reflect the participant's belief regarding how helpful the explanations are for them to understand the game and provide correct guidance to the robot team toward plans that are better suited to the scenarios and their value functions.

Given the satisfactory score, we formulate the overall generation as a hidden Markov model–based sequential generation process capable of adopting the temporal dynamics of the human user's explanation utility. More precisely, at each step, we first predefine a set of templates, each of which is accompanied by a combination of attributes, for example, isCounterfactual, hasTarget.; these templates provide the basis of an explanation and are filled in according to relevant slots. Next, the explainer determines the optimal syntax that matches the human's syntactic utility based on the satisfactory score; see the "Sequential explanation generation" section in Supplementary Methods for detailed computational flow.

One distinguished attribute to highlight is isRitualized, stemmed from the term "ontogenetic ritualization" in evolutionary anthropology literature. Conventionally, ritualization is referred to the evidence that early infants learn to communicate, especially in a symbolic manner, not based on imitation but rather on an individual learning process (*50*). Tomasello and Call (*51*) argue that such communicative behavior is a communicative signal that can be formed by two individuals shaping each other's behavior in repeated instances of interaction over time. Similar phenomena have also been observed and investigated in other primates, such as great apes (*52*). For example, many individual chimpanzees come to use a stylized "arm-raise" to indicate that they are about to hit the other and thus initiate play (*51*). In this way, a behavior that was not at first a communicative signal would become one over time. Inspired by this nonverbal behavior, the process of ontogenetic ritualization can also be formed during human-robot teaming, specifically when understanding and reacting to explanations. Intuitively, human speakers are reluctant to repeat similar messages that they have already conveyed before and would rather deliver a more concise version. To achieve this goal,

we explicitly define the "ritualized form" of explanation templates (table S1).

Human experiment details and demographics

The protocol for human study was reviewed and approved by the University of California, Los Angeles (UCLA) North Institutional Review Board (IRB), ID no. 20-001767. Human participants were recruited from University of California, San Diego (UCSD) undergraduate students taking psychology courses and the UCLA Department of Psychology participant pool. All participants provided written informed consent and were compensated with course credit for their participation. A total of 167 students completed the introduction phase and passed the familiarization test (56, 53, and 58 for the proposal-only, brief-explanation, and full-explanation groups, respectively). Nineteen participants were removed from the analysis for failing the attention check during the game play, resulting in 148 participants (49, 47, and 52 for the proposal-only, brief-explanation, and full-explanation groups, respectively) considered in the final results. In Fig. 4, we report results after the removal of outliers that are 1.5 interquartile range (IQR) below the 25th percentile or above the 75th percentile, resulting in 135 valid participants (45 participants per group). Before the game starts, all participants were assigned to one of the three explanation groups and given one of seven value functions randomly. The explanations for the brief-explanation and full-explanation groups were generated as described in the "Utility-aware explanation generation" section; see additional details of explanation generation and templates in Supplementary Methods. Participants in the proposal-only group did not have access to any explanations.

The experiment included three phases: introduction, familiarization, and game play. In the introduction phase, participants were presented with the context and rules of the Scout Exploration Game. Icons, scores, and UI in the game were explained to the participants with both text descriptions and video demonstrations. Because participants in different explanation groups saw different UI in the game, we guaranteed that the UI in the video demonstrations is consistent with the one in the actual game; video demonstrations for other groups were not presented, which ensures the between-participant design. In the familiarization stage, participants were tested with multiple-choice questions about their understanding of the game flow, rules, and the UI. Participants who correctly answered all questions proceeded to game play. Participants having at least one wrong answer were asked to review the introduction and retake the familiarization test. Participants who could not pass the familiarization test twice or took more than 20 min before starting the game play were removed from the study. During the game, we asked participants to provide feedback to the proposals and estimate the machine's internal states, such as the scouts' current value function and their qualitative trust of the XAI system. The scouts' value estimation questions were asked every two rounds of communication, and the trust questions were asked every five rounds. Figure 5 (C and D) shows the scouts' value estimation question and the trust question, respectively. Note that human judgments about the value estimation and trust were not used to adjust scouts' behavior; these additional measures were only used for evaluation purposes. Credits were awarded to participants regardless of their familiarization test results. The computational model used for scouts' value alignment was the same across all groups to attribute the difference in the performance of bidirectional value alignment to the lack or distinction of explanations. Participants in the proposal-only, brief-explanation, and full-explanation groups communicated with the scouts for 15.4, 16.4, and 16.2 rounds on average, with the SD of 3.2, 4.0, and 3.7 rounds, respectively. The average time of game play was 20.8, 22.5, and 32.3 min, with the SD of 4.5, 6.3, and 10.7 min for the proposal-only, brief-explanation, and full-explanation groups, respectively.

Statistics

To measure the value alignment performance, we use the Kendall τ coefficient to compare the goals' importance ranking in the target value with the ranking in the value estimation. The null hypothesis is that explanations yield the same value alignment across different groups, and therefore, no difference in the ranking statistics would be observed. To compare two sets of values (for example, values estimated by scouts versus true values known to human users), we first rank the task goals by their corresponding values and then calculate the Kendall's τ coefficient between the two rankings. Because games have various lengths due to different explanation formats, group values, and individual differences in participants, we normalize game progress as a percentage calculated by dividing the number of current iterations by the total number of iterations. For all three groups, we remove participants who fail attention checks and outliers whose alignment results are 1.5 IQR below the 25th percentile or above the 75th percentile at any game progresses. For statistical test, we perform paired t test to compare the degree of value alignment from the beginning of the game with the end of the game. In addition, we perform one-way ANOVA and paired t test to examine the group-wise difference of value alignment at different stages of the game. The t test we use assumes two-tailed independent samples. For all the statistical tests, we use significance level $\alpha = 0.05$ and rejection region $P \leq 0.05$.

SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scirobotics.abm4183
Supplementary Methods
Figs. S1 to S3
Table S1
Movies S1 and S2
MDAR Reproducibility Checklist
References (*53*, *54*)

REFERENCES AND NOTES

1. N. Wiener, Some moral and technical consequences of automation. *Science* **131**, 1355–1358 (1960).
2. R. Klimoski, S. Mohammed, Team mental model: Construct or metaphor? *J. Manage.* **20**, 403–437 (1994).
3. V. Groom, C. Nass, Can robots be teammates? Benchmarks in human–robot teams. *Interact. Stud.* **8**, 483–500 (2007).
4. S. H. Schwartz, *Advances in Experimental Social Psychology* (Elsevier, 1992), vol. 25, pp. 1–65.
5. W. B. Rouse, J. A. Cannon-Bowers, E. Salas, The role of mental models in team performance in complex systems. *IEEE Trans. Syst. Man Cybern.* **22**, 1296–1308 (1992).
6. J. MacMillan, E. E. Entin, D. Serfaty, Communication overhead: The hidden cost of team cognition, in *Team Cognition: Understanding the Factors That Drive Process and Performance*, E. Salas, S. M. Fiore, Eds. (American Psychological Association, 2004).
7. A. Butchibabu, C. Sparano-Huiban, L. Sonenberg, J. Shah, Implicit coordination strategies for effective team communication. *Hum. Factors* **58**, 595–610 (2016).
8. V. V. Unhelkar, S. Li, J. A. Shah, Decision-making for bidirectional communication in sequential human-robot collaborative tasks, in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 2020), pp. 329–341.
9. P. Abbeel, A. Y. Ng, Apprenticeship learning via inverse reinforcement learning, in *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)* (Association for Computing Machinery, 2004.

10. W. B. Knox, P. Stone, Interactively shaping agents via human reinforcement: The TAMER framework, in *Proceedings of the Fifth International Conference on Knowledge Capture* (Association for Computing Machinery, 2009).
11. S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, A. L. Thomaz, Policy shaping: Integrating human feedback with reinforcement learning, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2013).
12. M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, S.-C. Zhu, A tale of two explanations: Enhancing human trust by explaining robot behavior. *Sci. Robot.* **4**, aay4663 (2019).
13. M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Association for Computing Machinery, 2016).
14. H. Liu, Y. Zhang, W. Si, X. Xie, Y. Zhu, S.-C. Zhu, Interactive robot knowledge patching using augmented reality, in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2018).
15. Q. Zhang, X. Wang, Y. N. Wu, H. Zhou, S.-C. Zhu, Interpretable CNNs for object classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3416–3431 (2020).
16. Z. Zhang, Y. Zhu, S.-C. Zhu, Graph-based hierarchical knowledge representation for robot task transfer from virtual to physical world, in *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2021).
17. T. Chakraborti, S. Sreedharan, Y. Zhang, S. Kambhampati, Plan explanations as model reconciliation: Moving beyond explanation as soliloquy, in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)* (AAAI Press, 2017), pp. 156–163.
18. Z. Gong, Y. Zhang, Behavior explanation as intention signaling in human-robot teaming, in *Proceedings of International Symposium on Robot and Human Interactive Communication (RO-MAN)* (IEEE, 2018), pp. 1005–1011.
19. A. Tabrez, S. Agrawal, B. Hayes, Explanation-based reward coaching to improve human performance via reinforcement learning, in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 2019).
20. H. S. Huang, D. Held, P. Abbeel, A. Dragan, Enabling robots to communicate their objectives. *Autonom. Robots* **43**, 309–326 (2019).
21. T. Yuan, H. Liu. L. Fan, Z. Zhen, T. Gao, Y. Zhu, S.-C. Zhu, Joint inference of states, robot knowledge, and human (false-) beliefs, in *Proceedings of International Conference on Robotics and Automation (ICRA)* (IEEE, 2020).
22. X. Gao, R. Gong, Y. Zhao, S. Wang, T. Shu, S.-C. Zhu, Joint mind modeling for explanation generation in complex human-robot collaborative tasks, in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (IEEE, 2020), pp. 1119–1126.
23. S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin, 2019).
24. H. H. Clark, *Using Language* (Cambridge Univ. Press, 1996).
25. P. A. Samuelson, A note on the pure theory of consumer's behaviour. *Economica* **5**, 61 (1938).
26. B. A. Huberman, The ecology of computation, in *Digest of Papers. COMPCON Spring 89. Thirty-Fourth IEEE Computer Society International Conference: Intellectual Leverage* (IEEE, 1988), p. 362.
27. M. K. Ho, M. Littman, J. MacGlashan, F. Cushman, J. L. Austerweil, Showing versus doing: Teaching by demonstration, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2016).
28. H. P. Grice, in *Speech Acts* (Brill, 1975), pp. 41–58.
29. N. D. Goodman, A. Stuhlmüller, Knowledge and implicature: Modeling language understanding as social cognition. *Topics Cognit. Sci.* **5**, 173–184 (2013).
30. P. Shafto, N. D. Goodman, T. L. Griffiths, A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cogn. Psychol.* **71**, 55–89 (2014).
31. L. Yuan, D. Zhou, J. Shen, J. Gao, J. L. Chen, Q. Gu, Y. N. Wu, S.-C. Zhu, Iterative teacher-aware learning, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2021).
32. J. A. Simpson, Psychological foundations of trust. *Curr. Dir. Psychol. Sci.* **16**, 264–268 (2007).
33. M. Rheu, J. Y. Shin, W. Peng, J. Huh-Yoo, Systematic review: Trust-building factors and implications for conversational agent design. *Int. J. Human Comput. Interact.* **37**, 81–96 (2021).
34. M. Johnson, A. Vera, No AI is an island: The case for teaming intelligence. *AI Mag.* **40**, 16–28 (2019).
35. A. Barreto, W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. P,. van Hasselt, D. Silver, Successor features for transfer in reinforcement learning, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2017).
36. N. Wang, D. V. Pynadath, S. G. Hill, Trust calibration within a human-robot team: Comparing automatically generated explanations, in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 2016).
37. J. C. Licklider, R. W. Taylor, The computer as a communication device. *Sci. Technol.* **76**, 21–31 (1968).
38. L. B. Resnick, J. M. Levine, S. Behrend, *Socially Shared Cognition* (American Psychological Association, 1991).
39. S. Arora, P. Doshi, A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intel.* **297**, 103500 (2021).
40. D. Silver, J. Veness, Monte-Carlo planning in large POMDPs, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2010).
41. A. Rothe, B. M. Lake, T. M. Gureckis, Do people ask good questions? *Comput. Brain Behav.* **1**, 69–89 (2018).
42. M. Chen, S. Nikolaidis, H. Soh, D. Hsu, S. Srinivasa, Planning with trust for human-robot collaboration, in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 2018), pp. 307–315.
43. N. J. Smith, N. Goodman, M. Frank, Learning and using language via recursive pragmatic reasoning about other agents, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2013).
44. R. Carston, Informativeness, relevance and scalar implicature, in *Pragmatics And Beyond New Series* (John Benjamins Publishing Company, 1998), pp. 179–238.
45. A. Vogel, M. Bodoia, C. Potts, D. Jurafsky, Emergence of gricean maxims from multi-agent decision theory, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (ACM, 2013).
46. W. Liu, B. Dai, Z. Li, Z. Liu, J. Rehg, L. Song, Towards black-box iterative machine teaching, in *Proceedings of International Conference on Machine Learning (ICML)* (PMLR, 2018).
47. P. Wang, J. Wang, P. Paranamana, P. Shafto, A mathematical theory of cooperative communication, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2020).
48. H. de Weerd, R. Verbrugge, B. Verheij, Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonom. Agents Multi Agent Syst.* **31**, 250–287 (2017).
49. T. Peltola, M. M. Çelikok, P. Daee, S. Kaski, Machine teaching of active sequential learners, in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (IEEE, 2019).
50. A. Lock, *The Guided Reinvention of Language* (Academic Press, 1980).
51. M. Tomasello, J. Call, *Primate Cognition* (Oxford Univ. Press, 1997).
52. M. Tomasello, Do apes ape, in *Social Learning in Animals: The Roots of Culture*, C. M. Heyes, B. G. Galef, Jr. Eds. (Academic Press, 1996), pp. 319–346.
53. Z. Tu, S.-C. Zhu, Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 657 (2002).
54. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087 (1953).

Acknowledgments: The protocol for human study was reviewed and approved by the UCLA North IRB (ID no. 20-001767). **Funding:** This work was supported by DARPA XAI N66001-17-2-4029. **Author contributions:** L.Y.: devising algorithms, coding, designing and running participant study, analyzing data, and writing. X.G.: building the game interface, devising algorithms, coding, designing participant study, analyzing data, and writing. Z.Z.: devising algorithms, coding, designing participant study, analyzing data, and writing. M.E.: building the game interface, devising algorithms, coding, designing participant study, analyzing data, and writing. Y.N.W.: devising algorithms, writing, and providing the environment and the funding support for conducting this research. F.R.: designing and running participant study and writing. H.L.: designing participant study, examining data processing, and writing. Y.Z.: designing participant study, examining data processing, and writing. S.-C.Z.: setting research direction and providing the environment and the funding support for conducting this research. **Competing interests:** S.-C.Z. has affiliations with Beijing Institute for General Artificial Intelligence (BIGAI), Peking University, and Tsinghua University; Y.Z. is currently an employee of Peking University; M.E. is currently an employee of Cruise Automation; Z.Z. is currently an employee of BIGAI; X.G. has affiliations with Amazon Inc. The other authors declare that they have no competing interests. However, the research presented in this article is funded primarily by the DARPA XAI project and primarily conducted at UCLA and UCSD; later study analysis and paper writing are partially conducted while Y.Z., Z.Z., and S.-C.Z. are with Peking University and BIGAI. **Data and materials availability:** All data and software needed to evaluate the study of this paper are available in the paper or the Supplementary Materials. The code and data for this work have been deposited in the Dryad database https://doi.org/10.5068/D1XT3V.

Submitted 18 November 2021
Accepted 21 June 2022
Published 13 July 2022
10.1126/scirobotics.abm4183

Science Robotics

In situ bidirectional human-robot value alignment

Luyao YuanXiaofeng GaoZilong ZhengMark EdmondsYing Nian WuFederico RossanoHongjing LuYixin ZhuSong-Chun Zhu

Sci. Robot., 7 (68), eabm4183. • DOI: 10.1126/scirobotics.abm4183

Science Robotics (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.