HUMAN-ROBOT INTERACTION



# A tale of two explanations: Enhancing human trust by explaining robot behavior

**Mark Edmonds[1]*†, Feng Gao[2]*, Hangxin Liu[1]*, Xu Xie[2]*, Siyuan Qi[1], Brandon Rothrock[3], Yixin Zhu[2]†, Ying Nian Wu[2], Hongjing Lu[2,4], Song-Chun Zhu[1,2]†**

**The ability to provide comprehensive explanations of chosen actions is a hallmark of intelligence. Lack of this ability impedes the general acceptance of AI and robot systems in critical tasks. This paper examines what forms of explanations best foster human trust in machines and proposes a framework in which explanations are generated from both functional and mechanistic perspectives. The robot system learns from human demonstrations to open medicine bottles using (i) an embodied haptic prediction model to extract knowledge from sensory feedback, (ii) a stochastic grammar model induced to capture the compositional structure of a multistep task, and (iii) an improved Earley parsing algorithm to jointly leverage both the haptic and grammar models. The robot system not only shows the ability to learn from human demonstrators but also succeeds in opening new, unseen bottles. Using different forms of explanations generated by the robot system, we conducted a psychological experiment to examine what forms of explanations best foster human trust in the robot. We found that comprehensive and real-time visualizations of the robot's internal decisions were more effective in promoting human trust than explanations based on summary text descriptions. In addition, forms of explanation that are best suited to foster trust do not necessarily correspond to the model components contributing to the best task performance. This divergence shows a need for the robotics community to integrate model components to enhance both task execution and human trust in machines.**

## INTRODUCTION

Centuries ago, Aristotle stated that "we do not have knowledge of a thing until we have grasped its why, that is to say, its cause" (*1*). A hallmark of humans as social animals is the ability to answer this "why" question by providing comprehensive explanations of the behavior of themselves and others. The drive to seek explanations is deeply rooted in human cognition. Preschool-age children tend to attribute functions to all kinds of objects—clocks, lions, clouds, and trees—as explanations of the activity that these objects were apparently designed to perform (*2*, *3*). The strong human preference and intrinsic motivation for explanation are likely due to its central role in promoting mutual understanding, which fosters trust between agents and thereby enables sophisticated collaboration (*4*, *5*).

However, a strong human desire for explanations has not been sufficiently recognized by modern artificial intelligence (AI) systems, in which most methods primarily focus on task performance (*6*). Consequently, robot systems are still in their infancy in developing the ability to explain their own behavior when confronting noisy sensory inputs and executing complex multistep decision processes. Planner-based robot systems can generally provide an interpretable account for their actions to humans [e.g., by Markov decision processes (*7*, *8*), HTN (*9*), or STRIPS (*10*)], but these planners struggle to explain how their symbolic-level knowledge is derived from low-level sensory inputs. In contrast, robots equipped with deep neural networks (DNNs) (*11*) have demonstrated impressive performance in certain specific tasks due to their powerful ability to handle low-level noisy sensory inputs (*12*, *13*). However, DNN-based methods have well-known limitations, notably including a lack of interpretability of the knowledge representation (*14*–*16*). Some recent DNN work addressed this issue using saliency maps (*17*, *18*) or modularized components (*19*, *20*). These data-driven approaches have demonstrated strong capabilities of handling noisy real-time sensory inputs, distilling the raw input to predict the effect and determine the next action. However, little work has been done to develop the synergy between the classic symbolic AI and the recent development of DNNs to empower machines with the ability to provide comprehensive explanations of their behavior.

To fill in this gap, the present project aims to disentangle explainability from task performance, measuring each separately to gauge the advantages and limitations of two major families of representations—symbolic representations and data-driven representations—in both task performance and fostering human trust. The goals are to explore (i) what constitutes a good performer for a complex robot manipulation task? (ii) How can we construct an effective explainer to explain robot behavior and foster human trust?

To answer these questions, this paper develops an integrated framework consisting of a symbolic action planner using a stochastic grammar as the planner-based representation and a haptic prediction model based on neural networks to form the data-driven representation. We examined this integrated framework in a robot system using a contact-rich manipulation task of opening medicine bottles with various safety lock mechanisms. From the performer's perspective, this task is a challenging learning problem involving subtle manipulations, because it requires a robot to push or squeeze the bottle in various places to unlock the cap. At the same time, the task is also challenging for explanation, because visual information alone from a human demonstrator is insufficient to provide an effective explanation. Rather, the contact forces between the agent and the bottle provide the hidden "key" to unlock the bottle, and these forces cannot be observed directly from visual input.

[1]Department of Computer Science, UCLA, Los Angeles, CA 90095, USA. [2]Department of Statistics, UCLA, Los Angeles, CA 90095, USA. [3]Jet Propulsion Laboratory, Caltech, Los Angeles, CA 91109, USA. [4]Department of Psychology, UCLA, Los Angeles, CA 90095, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: markedmonds@ucla.edu (M.E.); yixin.zhu@ucla.edu (Y.Z.); sczhu@stat.ucla.edu (S.-C.Z.)

To constitute a good performer, the robot system proposed here cooperatively combined multiple sources of information for high performance, enabling synergy between a high-level symbolic action planner and a low-level haptic prediction model based on sensory inputs. A stochastic grammar model was learned from human demonstrations and served as a symbolic representation capturing the compositional nature and long-term constraints of a task (*21*). A haptic prediction model was trained using sensory information provided by human demonstrations (i.e., imposed forces and observed human poses) to acquire knowledge of the task. The symbolic planner and haptic model were combined in a principled manner using an improved generalized Earley parser (GEP) (*22*), which predicts the next robot action by integrating the high-level symbolic planner with the low-level haptic model. The learning from demonstration framework presented here shares a similar spirit of our previous work (*23*) but with a new haptic model and a more principled manner, namely, the GEP, to integrate the haptic and grammar models. Computational experiments demonstrate a strong performance improvement over the symbolic planner or the haptic model alone.

To construct an effective explainer, the proposed approach draws from major types of explanations in human learning and reasoning that may constitute representations to foster trust by promoting mutual understanding between agents. Previous studies have suggested that humans generate explanations from functional perspectives that describe the effects or goals of actions and from mechanistic perspectives that focus on behavior as a process (*24*). The haptic prediction model is able to provide a functional explanation by visualizing the essential haptic signals (i.e., effects of the previous action) to determine the next action. The symbolic action planner is capable of providing a mechanistic explanation by visualizing multiple planning steps (instead of just one) to describe the process of the task. The proposed robot system provides both functional and mechanistic explanations using the haptic model and the symbolic planner, respectively.

To examine how well robot-generated explanations foster human trust, we conducted human experiments to assess whether explanations provided by the robot system can foster trust in human users, and if so, what forms of explanation are the most effective in enhancing human trust in machines. Here, we refer to the cognitive component of "trust" (*25*) based on rationality. Cognitive trust is especially important in forming trust within secondary groups (such as human-machine relations) (*26*) compared with the emotional component typically more important in primary group relations (such as family and close friends). Our psychological experiment focuses on cognitive trust, stressing on a belief or an evaluation with "good rational reasons," because this is the crucial ingredient of human-machine trust built on specific beliefs and goals with attention to evaluations and expectations (*27*). Specifically, human participants were asked to report qualitative trust ratings after observing robot action sequences along with different forms of explanations for the robot's internal decision-making as it solved a manipulation task. Then, participants observed similar but new robot executions without access to explanations and were asked to predict how the robot system is likely to behave across time.

These empirical findings shed light on the importance of learning human-centric models that make the robot system explainable, trustworthy, and predictable to human users. Our results show that forms of explanation that are best suited to foster trust do not necessarily correspond to those components contributing to the best task performance. This divergence shows a need for the robotics community to adopt model components that are more likely to foster human trust and integrate these components with other model components enabling high task performance.

## RESULTS

Figure 1 illustrates the overall procedures, wherein the proposed integration framework, the GEP (*22*), efficiently combines a symbolic action planner and a data-driven haptic model to achieve high task performance and effective explanation. To this end, we first describe the procedure and data collection of human demonstrations, followed by the learning approaches. Next, we provide quantitative results as the success rate of the robot system in performing the task and assess the contributions from different modules of the system in task performance. We end the section with an analysis of human experiments with different types of explanations generated from the learned models, showing how human qualitative trust and prediction accuracy are influenced by various forms of explanations.

### Robot learning

To learn from human demonstrations, our robot system used an efficient encoding and representation of both haptic inputs and symbolic semantics of the manipulation task. The specific task, opening medicine bottles, requires inferring both the hand pose and the forces imposed on the bottle; agents must understand and enact the correct sequence of pose and force manipulations to succeed based on both the learned knowledge from human demonstrations and the real-time haptic sensory input.

We used a tactile glove with force sensors (*28*) to capture both the poses and the forces involved in human demonstrations in opening medicine bottles that require a visually latent interaction between the hand and the cap, e.g., pushing as indicated in Fig. 1A. A total of 64 human demonstrations, collected in (*23*), of opening three different medicine bottles served as the training data. These three bottles have different locking mechanisms: no safety lock mechanism, a push-twist locking mechanism, and a pinch-twist locking mechanism. To test the generalization ability of the robot system, we conducted a generalization experiment with new scenarios different from training data, either a new bottle (see "Robot results" section) or a bottle with a modified cap with significantly different haptic signals (fig. S1). The locking mechanisms of the bottles in the generalization experiment were similar but not identical (in terms of size, shape, and haptic signals) to the bottles used in human demonstrations. The haptic signals for the generalization bottles were significantly different from bottles used in testing, posing challenges in transferring the learned knowledge to novel unseen cases.

#### *Embodied haptic model*

Using human demonstrations, the robot learned a manipulation strategy based on the observed poses and forces exerted by human demonstrators. One challenge in learning manipulation policies from human demonstration involves different embodiments between robots and human demonstrators. A human hand has five fingers, whereas a robot gripper may only have two or three fingers; each embodiment exerts different sensory patterns even when performing the very same manipulation. Hence, the embodied haptic model for the robot system cannot simply duplicate human poses and forces exerted by human hands; instead, a robot should imitate the actions with the goal to produce the same end effect in manipulating the medicine bottle (e.g,, imposing a certain force on the cap). The critical approach in our model is to use embodied prediction, i.e., let the robot imagine

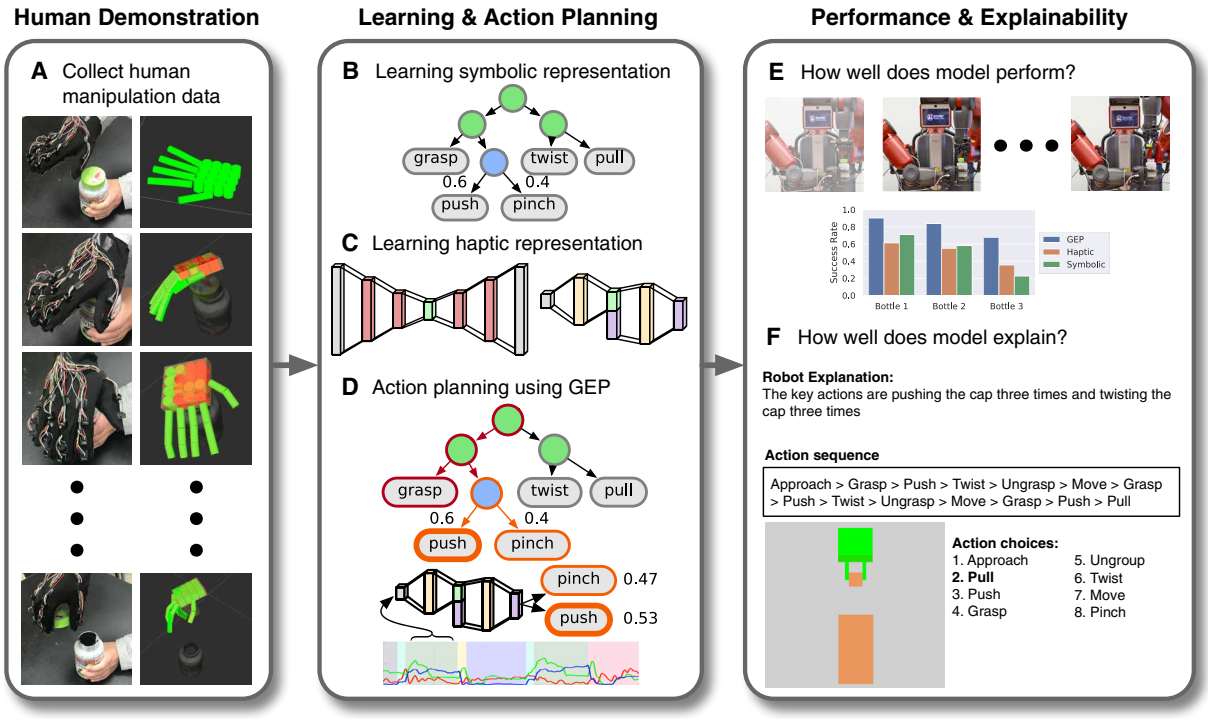



**Fig. 1. Overview of demonstration, learning, evaluation, and explainability.** By observing human demonstrations, the robot learns, performs, and explains using both a symbolic representation and a haptic representation. (**A**) Fine-grained human manipulation data were collected using a tactile glove. On the basis of the human demonstrations, the model learns (**B**) symbolic representations by inducing a grammar model that encodes long-term task structure to generate mechanistic explanations and (**C**) embodied haptic representations using an autoencoder to bridge the human and robot sensory input in a common space, providing a functional explanation of robot action. These two components are integrated using (**D**) the GEP for action planning. These processes complement each other in both (**E**) improving robot performance and (**F**) generating effective explanations that foster human trust.

its current haptic state as a human demonstrator and predict what action the demonstrator would have executed under similar circumstances in the next time step.

Figure 2 illustrates the force patterns exerted by a robot and a human demonstrator. As shown in Fig. 2 (A and C), due to the differences between a robot gripper and a human hand, the haptic sensing data from robots and humans show very different patterns from each other in terms of dimensionality and duration within each segmented action (illustrated by the colored segments).

To address the cross-embodiment problem, we trained a haptic model in a similar approach as in (*23*) to predict which action the robot should take next based on perceived human and robot forces and poses. The present haptic model learned a prediction model in a three-step process: (i) learning an autoencoder that constructs a low-dimensional embedding of human demonstrations containing poses and forces, as shown in Fig. 2B; (ii) training an embodiment mapping to map robot states to equivalent human embeddings, thereby allowing the robot to imagine itself as a human demonstrator to produce the same force, achieving functional equivalence to generate the same end effect as the human demonstrator (this embodiment mapping is trained in a supervised fashion, using labeled equivalent robot and human states); and (iii) training a next action predictor based on the human embeddings and the current action. This action predictor is also trained in a supervised fashion, using segmented human demonstrations (see the "Embodied haptic model details" section in Materials and Methods for additional training details).

The robot predicts the next action based on the mapped human embedding using a multi-class classifier (see details in the "Model learning details" section in Materials and Methods). We denote this prediction process as our haptic model. Intuitively, the embodied haptic predictions endow the robot with the ability to ask itself: "If I imagine myself as the human demonstrator, which action would the human have taken next based on the poses and forces exerted by their hand?" Hence, the resulting haptic model provides a functional explanation regarding the forces exerted by the robot's actions.

### *Symbolic action planner*

Opening medicine bottles is a challenging multistep manipulation, because one may need to push on the cap to unlock it (visually unobservable), twist it, and then pull it open. A symbolic representation is advantageous to capture the necessary long-term constraints of the task. From labeled action sequences of human demonstrations, we induce a temporal And-Or graph (T-AOG), a probabilistic graphical model describing a stochastic, hierarchical, and compositional context-free grammar (*29*), wherein an And-node encodes a decomposition of the graph into subgraphs, an Or-node reflects a switch among multiple alternate subconfigurations, and the terminal nodes consist of a set of action primitives (such as push, twist, and pull). A corpus of sentences (i.e., action sequences in our case) is fed to the grammar induction algorithm presented in (*21*), and the grammar is induced by greedily generating And-Or fragments according to the data likelihood; the fragments represent compositional substructures that are combined to form a complete grammar. In our case, the grammar was learned from segmented and labeled human demonstrations. The resulting grammar offers a compact symbolic representation of the task and captures the hierarchical structure of the task, including different action sequences for different

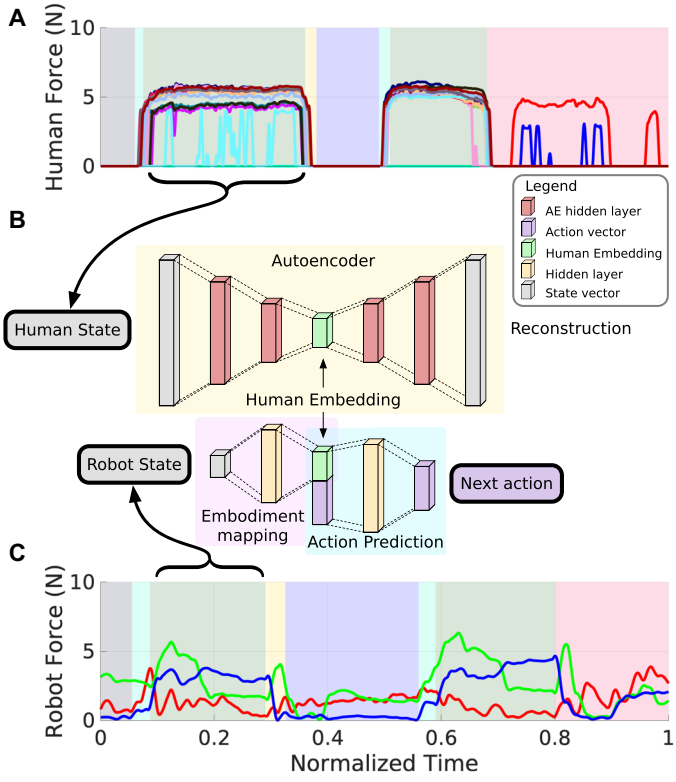



**Fig. 2. Illustration of learning embodied haptic representation and action prediction model.** An example of the force information in (**A**) the human state, collected by the tactile glove (with 26 dimensions of force data), and force information in (**C**) the robot state, recorded from the force sensors in the robot's end effector (with three dimensions of force data). The background colors indicate different action segments. For equivalent actions, the human and the robot may take different amounts of time to execute, resulting in different action segment lengths. (**B**) Embodied haptic representation and action prediction model. The autoencoder (yellow background) takes a human state, reduces its dimensionality to produce a human embedding, and uses the reconstruction to verify that the human embedding maintains the essential information of the human state. The embodiment mapping network (purple background) takes in a robot state and maps to an equivalent human embedding. The action prediction network (light blue background) takes the human embedding and the current action and predicts what action to take next. Thus, the robot imagines itself as a human based on its own haptic signals and predicts what action to take next.

bottles, as well as different action sequences for the same bottle. Examples of the T-AOG learning progress are shown in Fig. 3. The nodes connected by red edges in Fig. 3C indicate a parse graph sampled from the grammar, and its terminal nodes compose an action sequence for robot execution.

On the basis of the action sequences observed in human demonstrations, the induced grammar can be used to parse and predict robot action sequences that are likely to the successful opening of the medicine bottle, assuming each robot action corresponds to an equivalent human action. The induced grammar can be parsed to generate new, unseen, and valid action sequences for solving similar tasks (e.g., opening different medicine bottles), and thus, the grammar can be used with symbolic planning methods, such as the Earley parser (*22*). We denote the process of planning actions using a parser and the action grammar as the symbolic planner. Hence, the symbolic planner endows the robot with the ability to ask itself from a mechanistic perspective: "On the basis of what I have done thus far and what I observed the human do, which actions are likely to open the bottle at the end of the sequence?"

### *Integration of symbolic planner and haptic model*

To integrate the long-term task structure induced by the symbolic planner and the manipulation strategy learned from haptic signals, we sought to combine the symbolic action planner and embodied haptic model using the GEP (*22*). The GEP is a grammar parser that works on a sequence of sensory data; it combines any context-free grammar model with probabilistic beliefs over possible labels (grammar terminals) of sensory data. The output of the GEP is the optimal segmentation and label sentence of the raw sensory data; a label sentence is optimal when its probability is maximized according to the grammar priors and the input belief over labels while being grammatically correct. The core idea of the GEP is to efficiently search in the language space defined by the grammar to find the optimal label sentence.

To adopt the GEP for a robot system, we modified the GEP presented in (*22*) for online planning. The grammar for the GEP remains the same grammar used in the symbolic planner; however, the GEP's probabilistic beliefs come from the softmax distribution from the haptic model. During the action planning process, a stochastic distribution of action labels predicted by the haptic model is fed into the GEP at every time step. The GEP aggregates the entire symbolic planning history with the current haptic prediction and outputs the best parse to plan the most likely next action. Materials and Methods introduces more details about the algorithm. Intuitively, such an integration of the symbolic planner and haptic model enables the robot to ask itself: "On the basis of the human demonstration, the poses and forces I perceive right now, and the action sequence I have executed thus far, which action has the highest likelihood of opening the bottle?"

### *Robot results*

Figure 4 (A and B) shows the success rate of the robot opening the three medicine bottles used in human demonstrations and two new, unseen medicine bottles (see more generalization results in text S1). The two generalization bottles' locking mechanisms are similar (but not identical) to the ones used in human demonstrations, but the low-level haptic signals are significantly different, posing challenges in transferring the learned knowledge to novel unseen cases. To assess model performance, the robot attempted to open each bottle 31 times. In the testing experiments, bottle 1 is a regular bottle without a locking mechanism, bottle 2 has a push-twist locking mechanism, and bottle 3 requires pinching specific points on the lid to unlock. In the generalization experiments, bottle 4 also does not have a locking mechanism, and bottle 5 has a push-twist locking mechanism but with different shape, size, and haptic signals compared with the ones in the human demonstrations. For both the testing and generalization experiments, the robot's task performance measured by the success rates decreased as the bottle's locking mechanism became more complex, as expected.

To quantitatively compare the difference between the model components, we conducted ablative experiments on robot task performance using only the symbolic planner and only the haptic model (see Fig. 4). The haptic model and symbolic planner vary in their relative individual performance, but the combined planner using the GEP yields the best performance for all cases. Hence, integrating both the long-term task structure provided by the symbolic planner and the real-time sensory information provided by the haptic model yields the best robot performance. The symbolic planner provides long-term action planning and ensures that the robot executes an

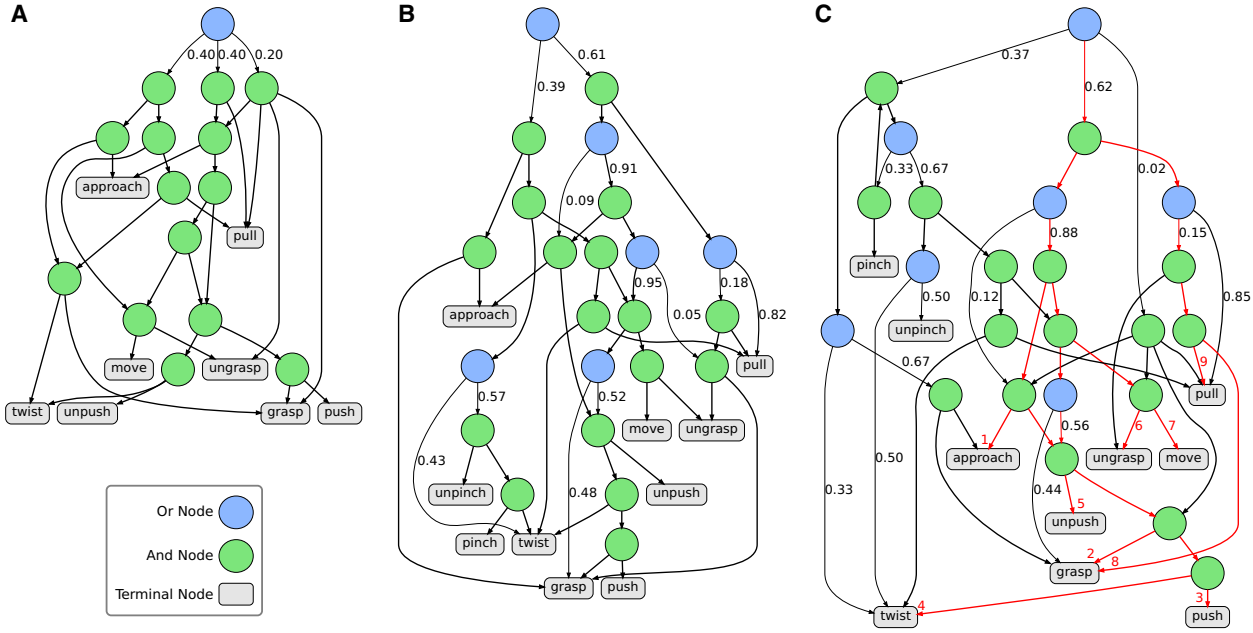



**Fig. 3. An example of action grammar induced from human demonstrations.** Green nodes represent And-nodes, and blue nodes represent Or-nodes. Probabilities along edges emanating from Or-nodes indicate the parsing probabilities of taking each branch. Grammar model induced from (**A**) 5 demonstrations, (**B**) 36 demonstrations, and (**C**) 64 demonstrations. The grammar model in (C) also shows a parse graph highlighted in red, where red numbers indicate temporal ordering of actions.

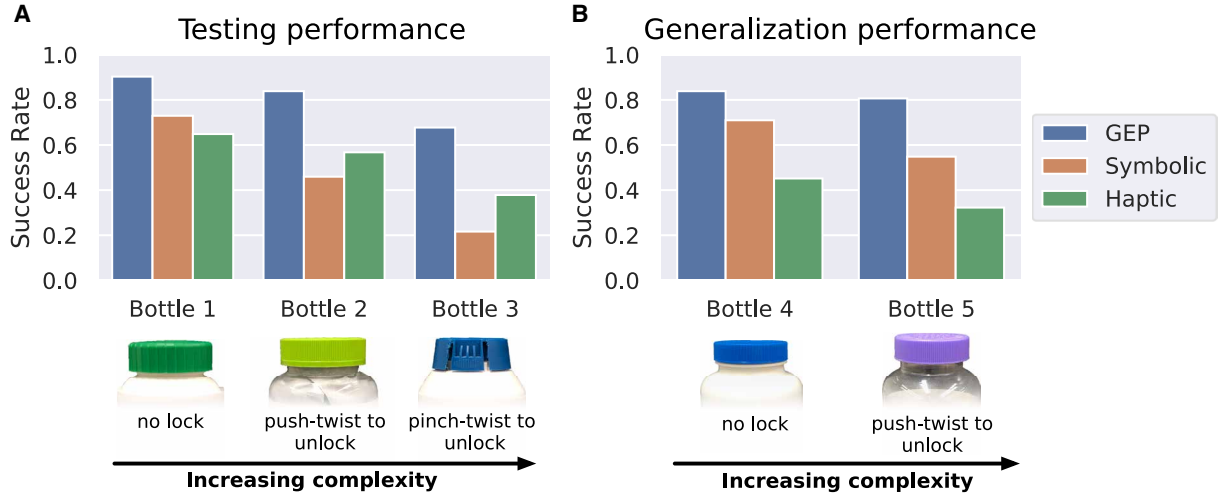



**Fig. 4. Robot task performance on different bottles with various locking mechanisms using the symbolic planner, haptic model, and the GEP that integrates both.** (**A**) Testing performance on bottles observed in human demonstrations. Bottle 1 does not have a locking mechanism, bottle 2 uses a push-twist locking mechanism, and bottle 3 uses a pinch-twist locking mechanism. (**B**) Generalization performance on new, unseen bottles. Bottle 4 does not have a locking mechanism, and bottle 5 uses a push-twist locking mechanism. The bottles used in generalization have similar locking mechanisms but evoke significantly different haptic feedback (see text S1). Regardless of testing on demonstration or unseen bottles, the best performance is achieved by the GEP that combines the symbolic planner and haptic model.

action sequence capturing the high-level structure of the task. However, models that solely rely on these symbolic structures are brittle to adjust to perturbations of haptic signals, especially when the task relies more on the haptics as the complexity increases. On the other hand, models that rely purely on haptic signals are unable to impose multistep task constraints and thus may fail to infer a correct sequence of actions based on the execution history. Our results confirm that by combining these modalities together, the robot achieves the highest task performance.

Given that multiple modalities are involved in the GEP's performance, it is crucial to assess the contributions from different model components. We ran the $\chi^2$ test to determine whether different models (GEP, symbolic, and haptic) are statistically different in their ability to open five bottles (three bottles used in human demonstrations and two new bottles used in the generalization task). The robot performed the manipulation task 31 times per medicine bottle. With the significance level of 0.05, the results show that the performance of the GEP model is significantly better than both symbolic

model [$\chi^2(1) = 10.0916$, $P = 0.0015$] and haptic model [$\chi^2(1) = 13.0106$, $P < 0.001$]. Performance does not show a difference between the symbolic model and the haptic model, $\chi^2(1) = 0.1263$, $P = 0.7232$. These results suggest that both haptic model and symbolic planner contribute to good task performance; when the two processes were integrated with the GEP, the success rate of the robot for opening medicine bottles was improved compared with the performance by the single-module models based on either the haptic model or the symbolic planner.

## Explanation generation

The haptic model and symbolic planner are capable of providing explanations to humans about robot behavior in real time. Mechanistic explanations can be generated by the symbolic planner in the form of action sequences because they represent the process of opening a medicine bottle. Functional explanations can be provided by a visualization of the internal robot gripper state (effects) used in the haptic model. It is worth noting that these models are capable of providing such explanations but are not the only means of producing them. Alternative action planners and haptic models could produce similar explanations, as long as the robot systems are able to learn the corresponding representations for haptic prediction and task structure. Figure 5 shows the explanation panels over an action sequence. These visualizations are shown in real time, providing direct temporal links between explanation and execution.

## Human experiment

### *Experimental design*

The human experiment aims to examine whether providing explanations generated from the robot's internal decisions fosters human trust to machines and what forms of explanation are the most effective in enhancing human trust. We conducted a psychological study with 150 participants; each was randomly assigned to one of five groups. Our experimental setup consisted of two phases: familiarization and prediction. During familiarization, all groups viewed the RGB video, and some groups were also provided with an explanation panel. During the second phase of the prediction task, all groups only observed RGB videos.

The five groups consist of the baseline no-explanation group, symbolic explanation group, haptic explanation group, GEP explanation group, and text explanation group. For the baseline no-explanation group, participants only viewed RGB videos recorded from a robot attempting to open a medicine bottle, as shown in Fig. 6A. For the other four groups, participants viewed the same RGB video of robot executions and simultaneously were presented with different explanatory panels on the right side of the screen.

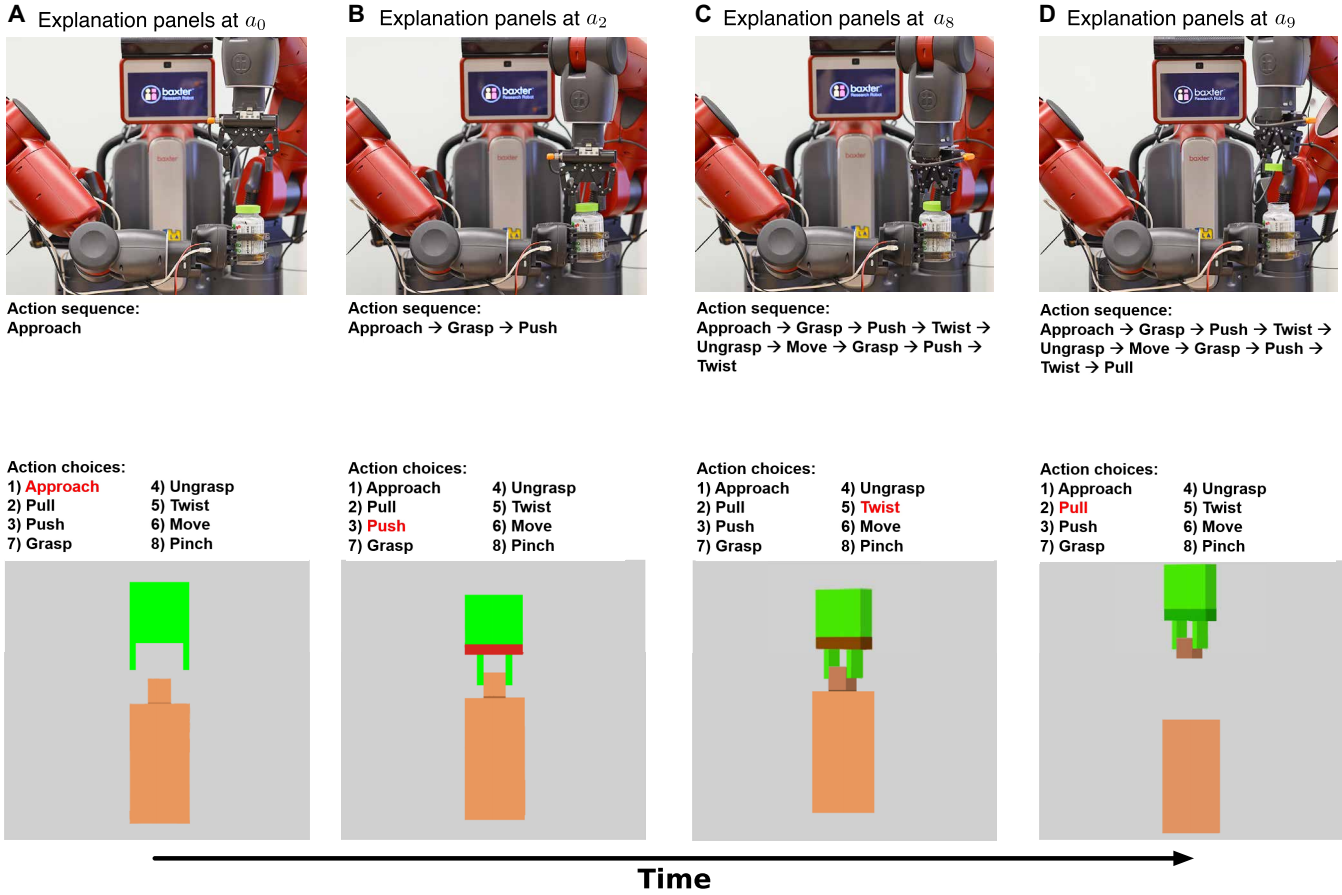
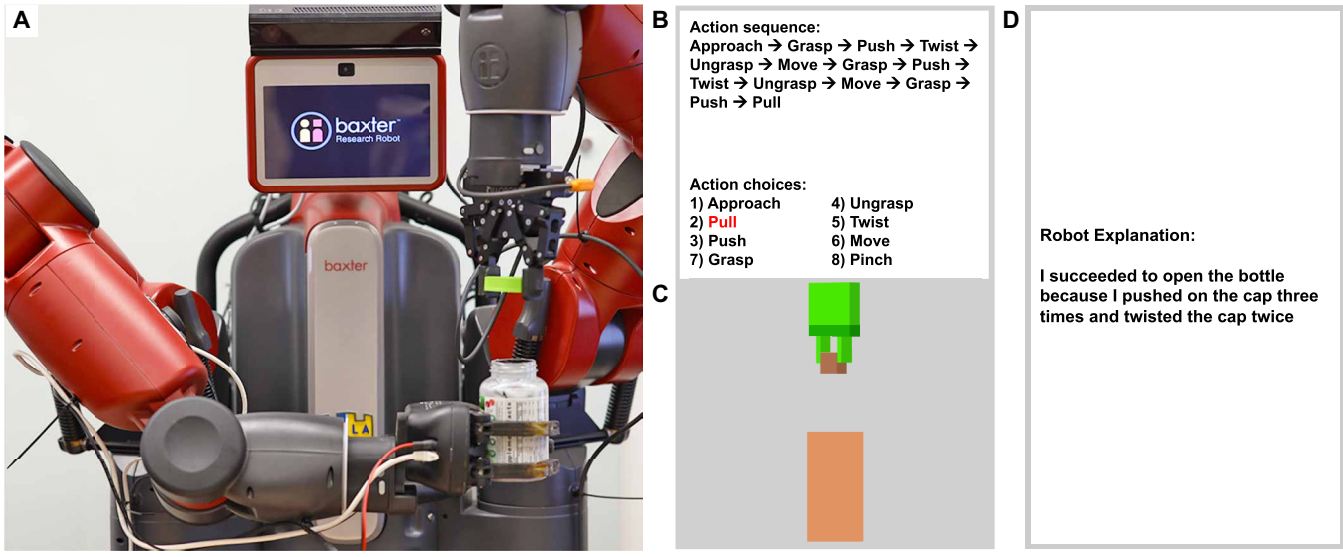



**Fig. 5. Explanations generated by the symbolic planner and the haptic model.** (**A**) Symbolic (mechanistic) and haptic (functional) explanations at $a_0$ of the robot action sequence. (**B** to **D**) Explanations at times $a_2$, $a_8$, and $a_9$, respectively, where $a_i$ refers to the $i$th action. Note that the red on the robot gripper's palm indicates a large magnitude of force applied by the gripper, and green indicates no force; other values are interpolated. These explanations are provided in real time as the robot executes.




**E** Summary of human subject groups and explanations presented

| Group | RGB (A) | Symbolic (B) | Haptic (C) | Text (D) |
|---|---|---|---|---|
| Baseline | ✓ | | | |
| Symbolic | ✓ | ✓ | | |
| Haptic | ✓ | | ✓ | |
| GEP | ✓ | ✓ | ✓ | |
| Text | ✓ | | | ✓ |

**Fig. 6. Illustration of visual stimuli used in human experiment.** All five groups observed the RGB video recorded from robot executions but differed by the access to various explanation panels. (**A**) RGB video recorded from robot executions. (**B**) Symbolic explanation panel. (**C**) Haptic explanation panel. (**D**) Text explanation panel. (**E**) Summary of which explanation panels were presented to each group.

Specifically, the symbolic group viewed the symbolic action planner illustrating the robot's real-time inner decision-making, as shown in Fig. 6B. The haptic group viewed the real-time haptic visualization panel, as shown in Fig. 6C. The GEP group viewed the combined explanatory panel, including the real-time robot's symbolic planning and an illustration of haptic signals from the robot's manipulator, namely, both Fig. 6, B and C. The text explanation group was provided a text description that summarizes why the robot succeeded or failed to open the medicine bottle at the end of the video, as shown in Fig. 6D (see a summary in Fig. 6E for the five experimental groups).

During the familiarization phase, participants were provided two demonstrations of robot executions, with one successful execution of opening a medicine bottle and one failed execution without opening the same bottle. The presentation order of the two demonstrations was counterbalanced across participants. After observing robot executions with explanation panels, participants were first asked to provide a trust rating for the question: "To what extent do you trust/believe this robot has the ability to open a medicine bottle, on a scale between 0 and 100?" The question was adopted from the questionnaire of measuring human trust in automated systems (*30*). This question also clearly included the goal of the system, to open a medicine bottle, to enhance the reliability in trust measures (*27*). Hence, the rating provided a direct qualitative measure of human trust to the robot's ability to open medicine bottles.

In addition, we designed the second measure to assess the quantitative aspects of trust. We adopted the definition by Castelfranchi and Falcone (*27*) that quantitative trust is based on the quantitative dimensions of its cognitive constituents. Specifically, the greater the human's belief in the machine's competence and performance, the greater the human trust in machines. In the prediction phase, we asked participants to predict the robot's next action in a new execution with the same task of opening the same medicine bottle. Participants viewed different segments of actions performed by the robot and were asked to answer the prediction question over time. For this measure, participants in all five groups only observed RGB videos of robot execution during the prediction phase; no group had access to any explanatory panel after the familiarization phase. The prediction accuracy was computed as the quantitative measure of trust, with the presumption that, as the robot behavior is more predictable to humans, greater prediction accuracy indicates higher degrees of trust.

### *Human study results*

Figure 7A shows human trust ratings from the five different groups. The analysis of variance (ANOVA) reveals a significant main effect of groups ($F_{4,145} = 2.848$; $P = 0.026$) with the significance level of 0.05. This result suggests that providing explanations about robot behavior in different forms affects the degree of human trust to the robot system. Furthermore, we found that the GEP group with both symbolic and haptic explanation panels yields the highest trust rating, with a significantly better rating than the baseline group in which explanations are not provided [independent-samples $t$ test, $t(58) = 2.421$; $P = 0.019$]. The GEP group showed a greater trust rating than the text group in which a summary description was provided to explain the robot behavior [$t(58) = 2.352$; $P = 0.022$], indicating that detailed explanations of robot's internal decisions over time is much more

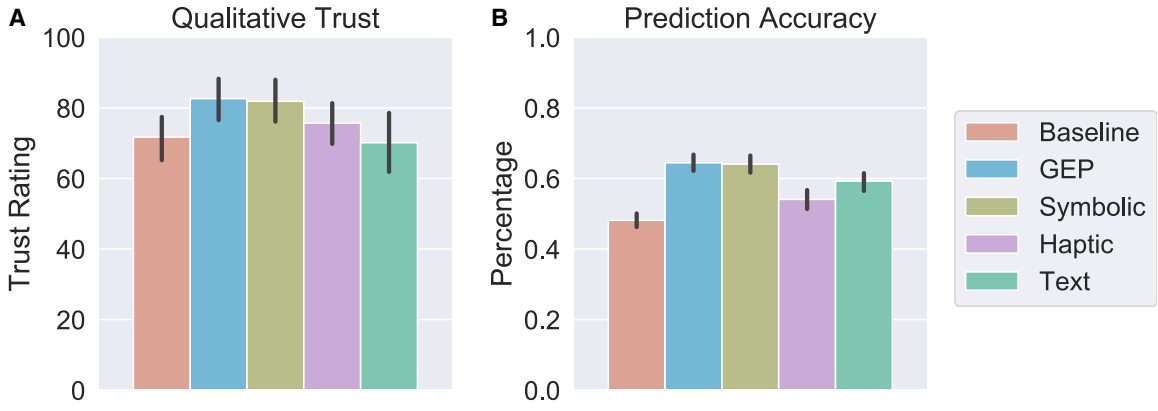



**Fig. 7. Human results for trust ratings and prediction accuracy.** (**A**) Qualitative measures of trust: average trust ratings for the five groups. (**B**) Average prediction accuracy for the five groups. The error bars indicate the 95% confidence interval. Across both measures, the GEP performs the best. For qualitative trust, the text group performs most similarly to the baseline group. For a tabular summary of the data, see table S1.

effective in fostering human trust than a summary text description to explain robot behavior. In addition, trust ratings in the symbolic group were also higher than ratings in the baseline group [$t(58) = 2.269$; $P = 0.027$] and higher than ratings in the text explanation group [$t(58) = 2.222$; $P = 0.030$], suggesting that symbolic explanations play an important role in fostering human trust of the robot system. However, the trust ratings in the haptic explanation group were not significantly different from the baseline group, implying that explanations only based on haptic signals are not effective ways to gain human trust despite the explanations also being provided in real time. No other significant group differences were observed between any other pairing of the groups.

The second trust measure based on prediction accuracy yielded similar results. All groups provided action predictions above the chance-level performance of 0.125 (as there are eight actions to choose from), showing that humans are able to predict the robot's behavior after only a couple of observations of a robot performing a task. The ANOVA analysis shows a significant main effect of groups ($F_{4,145} = 3.123$; $P = 0.017$), revealing the impact of provided explanations on the accuracy of predicting the robot's actions. As shown in Fig. 7B, participants in the GEP group yielded significantly higher prediction accuracy than those in the baseline group [$t(58) = 3.285$; $P = 0.002$]. Prediction accuracy of the symbolic group also yielded better performance than the baseline group [$t(58) = 2.99$; $P = 0.004$]. We found that the text group shows higher prediction accuracy than the baseline group [$t(58) = 2.144$; $P = 0.036$]. This result is likely due to the summary text description providing a loose description of the robot's action plan; such a description decouples the explanation from the temporal execution of the robot. The prediction accuracy data did not reveal any other significant group differences among other pairs of groups.

In general, humans appear to need real-time, symbolic explanations of the robot's internal decisions for performed action sequences to establish trust in machines performing multistep complex tasks. Summary text explanations and explanations only based on haptic signals are not effective ways to gain human trust, and the GEP and symbolic group foster similar degrees of human trust to the robot system according to both measures of trust.

## DISCUSSION

In terms of performance, our results demonstrate that a robot system can learn to solve challenging tasks from a small number of human demonstrations of opening three medicine bottles. This success in learning from small data is primarily supported by learning multiple models for joint inference of task structure and sensory predictions. We found that neither purely symbolic planning nor a haptic model is as successful as an integrated model including both processes.

Our model results also suggest that the relative contributions from individual modules, namely, the symbolic planner and haptic predictions, can be influenced by the complexity of the manipulation task. For example, in testing scenarios, for bottle 1 with no safety locking mechanism, the symbolic planner slightly outperformed the haptic model. Conversely, to open bottle 3 that has complex locking mechanisms, the haptic model outperformed the symbolic planner because haptic signals provide critical information for the pinch action needed to unlock the safety cap. For generalization scenarios with new medicine bottles that were not seen in human demonstrations, the symbolic planner maintained a similar performance compared with equivalent bottles in the testing scenarios, whereas the haptic model performance decreased significantly. We also note that the symbolic planner performance decreased faster as complexity increased, indicating that pure symbolic planners are more brittle to circumstances that require additional haptic sensing. Furthermore, as bottle complexity increased, model performance benefitted more from integrating symbolic planner and haptic signals. This trend suggests that more complex tasks require the optimal combination of multiple models to produce effective action sequences.

In terms of explainability, we found that reasonable explanations generated by the robot system are important for fostering human trust in machines. Our experiments show that human users place more trust in a robot system that has the ability to provide explanations using symbolic planning. An intriguing finding from these experiments is that providing explanations in the form of a summarized text description of robot behavior is not an effective way to foster human trust. The symbolic explanation panel and text summary panel both provided critical descriptions of the robot's behavior at the abstract level, explaining why a robot succeeded or failed the task. However, the explanations provided by the two panels differed in their degree of detail and temporal presentation. The text explanation provided a loose description of the important actions that the robot executes after the robot finished the sequence. In contrast, the symbolic explanation included in the GEP's panel provided human participants with real-time visualizations of the robot's internal planning process

at each step. This mode of explanation enables the visualization of task structure for every action executed during the sequence and likely evokes a sense that the robot actively makes rational decisions.

However, it is not the case that a detailed explanation is always the best approach to foster human trust. A functional explanation of real-time haptic signals is not very effective in gaining human trust in this particular task. Information at the haptic level may be excessively tedious and may not yield a sense of rational agency that allows the robot to gain human trust. To establish human trust in machines and enable humans to predict robot behaviors, it appears that an effective explanation should provide a symbolic interpretation and maintain a tight temporal coupling between the explanation and the robot's immediate behavior.

Taking together both performance and explanation, we found that the relative contributions of different model components for generating explanations may differ from their contributions to maximizing robot performance. For task performance, the haptic model plays an important role for the robot to successfully open a medicine bottle with high complexity. However, the major contribution to gain human trust is made by real-time mechanistic explanations provided by the symbolic planner. Hence, model components that foster the most trust do not necessarily correspond to those components contributing to the best task performance. This divergence is possible because there is no requirement that components responsible for generating better explanations are the same components contributing to task performance; they are optimizing different goals. This divergence also implies that the robotics community should adopt model components that gain human trust while also integrating these components with high-performance components to maximize both human trust and successful execution. Robots endowed with explainable models offer an important step toward integrating robots into daily life and work.

## MATERIALS AND METHODS

### Embodied haptic model details

The embodied haptic model leverages low-level haptic signals obtained from the robot's manipulator to make action predictions based on the human poses and forces collected with the tactile glove. This embodied haptic sensing allows the robot to reason about (i) its own haptic feedback by imagining itself as a human demonstrator and (ii) what a human would have done under similar poses and forces. The critical challenge here is to learn a mapping between equivalent robot and human states, which is difficult due to the different embodiments. From the perspective of generalization, manually designed embodiment mappings are not desirable. To learn from human demonstrations on arbitrary robot embodiments, we propose an embodied haptic model general enough to learn between an arbitrary robot embodiment and a human demonstrator.

The embodied haptic model consists of three major components: (i) an autoencoder to encode the human demonstration in a low-dimensional subspace (we refer to the reduced embedding as the human embedding); (ii) an embodiment mapping that maps robot states onto a corresponding human embedding, providing the robot with the ability to imagine itself as a human demonstrator; and (iii) an action predictor that takes a human embedding and the current action executing as the input and predicts the next action to execute, trained using the action labels from human demonstrations. Figure 2B shows the embodied haptic network architecture. Using this network architecture, the robot infers what action a human was likely to execute on the basis of this inferred human state. This embodied action prediction model picks the next action according to

$$a_{t+1} \sim p(\cdot|f_t, a_t) \quad (1)$$

where $a_{t+1}$ is the next action, $f_t$ is the robot's current haptic sensing, and $a_t$ is the current action.

The autoencoder network takes an 80-dimensional vector from the human demonstration (26 for the force sensors and 54 for the poses of each link in the human hand) and uses the post-condition vector, i.e., the average of last $N$ frames (we choose $N = 2$ to minimize the variance), of each action in the demonstration as input (see the autoencoder portion of Fig. 2B). This input is then reduced to an eight-dimensional human embedding. Given a human demonstration, the autoencoder enables the dimensionality reduction to an eight-dimensional representation.

The embodiment mapping maps from the robot's four-dimensional post-condition vector, i.e., the average of the last $N$ frames (different from human post-condition due to a faster sample rate on the robot gripper compared with the tactile glove; we chose $N = 10$), to an imagined human embedding (see the embodiment mapping portion of Fig. 2B). This mapping allows the robot to imagine its current haptic state as an equivalent low-dimensional human embedding. The robot's four-dimensional post-condition vector consists of the gripper position (one dimension) and the forces applied by the gripper (three dimensions). The embodiment mapping network uses a 256-dimensional latent representation, and this latent representation is then mapped to the eight-dimensional human embedding.

To train the embodiment mapping network, the robot first executes a series of supervised actions where, if the action produces the correct final state of the action, the robot post-condition vector is saved as input for network training. Next, human demonstrations of equivalent actions are fed through the autoencoder to produce a set of human embeddings. These human embeddings are considered as the ground-truth target outputs for the embodiment mapping network, regardless of the current reconstruction accuracy of the autoencoder network. Then, the robot execution data are fed into the embodiment mapping network, producing an imagined human embodiment. The embodiment mapping network optimizes to reduce the loss between its output from the robot post-condition input and the target output.

For the action predictor, the 8-dimensional human embedding and the 10-dimensional current action are mapped to a 128-dimensional latent representation, and the latent representation is then mapped to a final 10-dimensional action probability vector (i.e., the next action) (see action prediction portion of Fig. 2B). This network is trained using human demonstration data, where a demonstration is fed through the autoencoder to produce a human embedding, and that human embedding and the one-hot vector of the current action execution are fed as the input to the prediction network; the ground truth is the next action executed in the human demonstration.

The network in Fig. 2B is trained in an end-to-end fashion with three different loss functions in a two-step process: (i) a forward pass through the autoencoder to update the human embedding $z_h$. After computing the error $L_{\text{reconstruct}}$ between the reconstruction $s'_h$ and the ground-truth human data $s_h$, we back-propagate the gradient and optimize the autoencoder

$$L_{\text{reconstruct}}(s_h', s_h) = \frac{1}{2}(s_h' - s_h)^2 \quad (2)$$

(ii) A forward pass through the embodiment mapping and the action prediction network. The embodiment mapping is trained by minimizing the difference $L_{\text{mapping}}$ between the embodied robot embedding $z_r$ and target human embedding $z_h$; the target human embedding $z_h$ is acquired through a forward pass through the autoencoder using a human demonstration post-condition of the same action label, $s_h$. We compute the cross-entropy loss $L_{\text{prediction}}$ of the predicted action label $a'$ and the ground-truth action label $a$ to optimize this forward pass

$$\begin{aligned} L_{\text{planning}}(a', a) &= L_{\text{mapping}} + \beta \cdot L_{\text{prediction}} \\ L_{\text{mapping}} &= \frac{1}{2}(z_r - z_h)^2 \\ L_{\text{prediction}} &= H(p(a'), q(a)) \end{aligned} \quad (3)$$

where $H$ is the cross entropy, $p$ is the model prediction distribution, $q$ is the ground-truth distribution, and β is the balancing parameter between the two losses (see text S2.2 for detailed parameters and network architecture).

A similar embodied haptic model was presented in (*23*) but with two separate loss functions, which is more difficult to train compared with the single loss function presented here. A clear limitation of the haptic model is the lack of long-term action planning. To address this problem, we discuss the symbolic task planner below and then discuss how we integrated the haptic model with the symbolic planner to jointly find the optimal action.

### Symbolic planner details

To encode the long-term temporal structure of the task, we endow a symbolic action planner that encodes semantic knowledge of the task execution sequence. The symbolic planner uses stochastic context-free grammars to represent tasks, where the terminal nodes (words) are actions and sentences are action sequences. Given an action grammar, the planner finds the optimal action to execute next on the basis of the action history, analogous to predicting the next word given a partial sentence.

The action grammar is induced using labeled human demonstrations, and we assume that the robot has an equivalent action for each human action. Each demonstration forms a sentence, $x_i$, and the collection of sentences from a corpus, $x_i \in X$. The segmented demonstrations are used to induce a stochastic context-free grammar using the method presented in (*21*). This method pursues T-AOG fragments to maximize the likelihood of the grammar producing the given corpus. The objective function is the posterior probability of the grammar given the training data $X$

$$p(G \mid X) \propto p(G)\, p(X \mid G) = \frac{1}{Z} e^{-\alpha \|G\|} \prod_{x_i \in X} p(x_i \mid G) \quad (4)$$

where $G$ is the grammar, $x_i = (a_1, a_2, \ldots, a_m) \in X$ represents a valid sequence of actions with length $m$ from the demonstrator, α is a constant, $\|G\|$ is the size of the grammar, and $Z$ is the normalizing factor. Figure 3 shows examples of induced grammars of actions.

During the symbolic planning process, this grammar is used to compute which action is the most likely to open the bottle based on the action sequence executed thus far and the space of possible future actions. A pure symbolic planner picks the optimal action based on the grammar prior

$$a_{t+1}^* = \arg\max_{a_{t+1}} p(a_{t+1} \mid a_{0:t}, G) \quad (5)$$

where $a_{t+1}$ is the next action and $a_{0:t}$ is the action sequence executed thus far. This grammar prior can be obtained by a division of two grammar prefix probabilities: $p(a_{t+1} \mid a_{0:t}, G) = \frac{p(a_{0:t+1} \mid G)}{p(a_{0:t} \mid G)}$, where the grammar prefix probability $p(a_{0:t} \mid G)$ measures the probability that $a_{0:t}$ occurs as a prefix of an action sequence generated by the action grammar $G$. On the basis of a classic parsing algorithm—the Earley parser (*31*)—and dynamic programming, the grammar prefix probability can be obtained efficiently by the Earley-Stolcke parsing algorithm (*32*). An example of pure symbolic planning is shown in fig. S4.

However, due to the fixed structure and probabilities encoded in the grammar, always choosing the action sequence with the highest grammar prior is problematic because it provides no flexibility. An alternative pure symbolic planner picks the next action to execute by sampling from the grammar prior

$$a_{t+1} \sim p(\cdot \mid a_{0:t}, G) \quad (6)$$

In this way, the symbolic planner samples different grammatically correct action sequences and increases the adaptability of the symbolic planner. In the experiments, we choose to sample action sequences from the grammar prior.

In contrast to the haptic model, this symbolic planner lacks the adaptability to real-time sensor data. However, this planner encodes long-term temporal constraints that are missing from the haptic model, because only grammatically correct sentences have nonzero probabilities. The GEP adopted in this paper naturally combines the benefits of both the haptic model and the symbolic planner (see the next section).

### GEP details

The robot imitates the human demonstrator by combining the symbolic planner and the haptic model. The integrated model finds the next optimal action considering both the action grammar $G$ and the haptic input $f_t$

$$a_{t+1}^* = \arg\max_{a_{t+1}} p(a_{t+1} \mid a_{0:t}, f_t, G) \quad (7)$$

Conceptually, this can be thought of as a posterior probability that considers both the grammar prior and the haptic signal likelihood. The next optimal action is computed by an improved GEP (*22*); GEP is an extension of the classic Earley parser (*31*). In the present work, we further extend the original GEP to make it applicable to multisensory inputs and provide explanation in real time for robot systems, instead of for offline video processing (see details in text S4.1.3).

The computational process of GEP is to find the optimal label sentence according to both a grammar and a classifier output of probabilities of labels for each time step. In our case, the labels are actions, and the classifier output is given by the haptic model. Optimality here means maximizing the joint probability of the action sequence according to the grammar prior and haptic model output while being grammatically correct.

The core idea of the algorithm is to directly and efficiently search for the optimal label sentence in the language defined by the grammar. The grammar constrains the search space to ensure that the sentence is always grammatically correct. Specifically, a heuristic search is performed on the prefix tree expanded according to the grammar, where the path from the root to a node represents a partial sentence (prefix of an action sequence).

GEP is a grammar parser, capable of combining the symbolic planner with low-level sensory input (haptic signals in this paper). The search process in the GEP starts from the root node of the prefix tree, which is an empty terminal symbol indicating that no terminals are parsed. The search terminates when it reaches a leaf node. In the prefix tree, all leaf nodes are parsing terminals *e* that represent the end of parse, and all non-leaf nodes represent terminal symbols (i.e., actions). The probability of expanding a non-leaf node is the prefix probability, i.e., how likely is the current path being the prefix of the action sequence. The probability of reaching a leaf node (parsing terminal *e*) is the parsing probability, i.e., how likely is the current path to the last non-leaf node being the executed actions and the next action. In other words, the parsing probability measures the probability that the last non-leaf node in the path will be the next action to execute. It is important to note that this prefix probability is computed on the basis of both the grammar prior and the haptic prediction; in contrast, in the pure symbolic planner, the prefix probability is computed on the basis of only the grammar prior. An example of the computed prefix and parsing probabilities and output of GEP is given by Fig. 8, and the search process is illustrated in fig. S5. For an algorithmic description of this process, see algorithm S1.

The original GEP is designed for offline video processing. Here, we made modifications to enable online planning for a robotic task. The major difference between parsing and planning is the uncertainty about past actions: There is uncertainty about observed actions during parsing. However, during planning, there is no uncertainty about executed actions—the robot directly chooses which actions to execute, thereby removing any ambiguity regarding which action was executed at a previous time step. Hence, we need to prune the impossible parsing results after executing each action; each time after executing an action, we change the probability vector of that action to a one-hot vector. This modification effectively prunes the action sequences that contain the impossible actions executed thus far by the robot.

## Tactile glove

For manipulation tasks that require reasoning about latent forces, demonstrations that contain solely visual information (e.g., RGB videos) are insufficient for learning. Using a glove-based system to capture hand-related data has long been proposed; however, it remains an active research topic due to the high articulation and degrees of freedom of a human hand. Conventionally, a network of inertial measurement units (IMUs) measures finger poses, but capturing haptic signals is challenging due to hand deformation and a scarcity of force sensing hardware. Here, we used the tactile glove developed in (*28*). The glove used IMUs to obtain the relative poses of finger phalanges with respect to the wrist and developed a customized force sensor using a soft piezoresistive material (Velostat) whose resistance changes under pressure (see more hardware details in text S2.3).

## Robot platform

We evaluate the learned model on a dual-armed seven-DoF Baxter robot mounted on a DataSpeed mobility base. The robot was equipped with a ReFlex TakkTile gripper on the right wrist and a Robotiq S85 parallel gripper on the left. The grippers have minimal haptic sensing capability; they can only determine whether or not the gripper is in contact with an object. Therefore, further force data on the robot were obtained from the 6 degrees-of-freedom (DOF) force and torque sensors located in Baxter's wrists. In addition, a Kinect One sensor was used for object pose estimation and tracking. The entire system ran on Robot Operating System, and the arm motion was planned by MoveIt!

## Human experiment details and demographics

Human participants were recruited from the University of California, Los Angeles (UCLA) Department of Psychology subject pool and were compensated with course credit for their participation. A total of 163 students were recruited, each randomly assigned to one of the five experimental groups. Thirteen participants were removed

**A** Input probability matrix

| Time step | grasp | push | pinch | twist | pull |
|---|---|---|---|---|---|
| 0 | 0.6 | 0.1 | 0.1 | 0.1 | 0.1 |
| 1 | 0.6 | 0.1 | 0.1 | 0.1 | 0.1 |
| 2 | 0.1 | 0.1 | 0.6 | 0.1 | 0.1 |
| 3 | 0.1 | 0.1 | 0.6 | 0.1 | 0.1 |
| 4 | 0.1 | 0.1 | 0.1 | 0.6 | 0.1 |
| 5 | 0.1 | 0.1 | 0.1 | 0.6 | 0.1 |

**B** Cached probabilities

| Time step | $\epsilon$ | grasp | grasp push | grasp pinch | grasp push twist | grasp pinch twist | grasp push twist pull | grasp pinch twist pull |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.000 | 0.600 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 1 | 0.000 | 0.360 | 0.036 | 0.024 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 0.000 | 0.036 | 0.025 | 0.101 | 0.004 | 0.002 | 0.000 | 0.000 |
| 3 | 0.000 | 0.004 | 0.005 | 0.069 | 0.003 | 0.010 | 3.6e-04 | 2.4e-04 |
| 4 | 0.000 | 3.6e-04 | 6.8e-04 | 0.007 | 0.005 | 0.048 | 3.2e-04 | 0.001 |
| 5 | 0.000 | 3.6e-05 | 9.0e-05 | 7.2e-04 | 0.003 | 0.033 | 4.9e-04 | 0.005 |
| prefix | 1.000 | 0.600 | 0.060 | 0.119 | 0.009 | 0.058 | 0.001 | 0.006 |

**C**

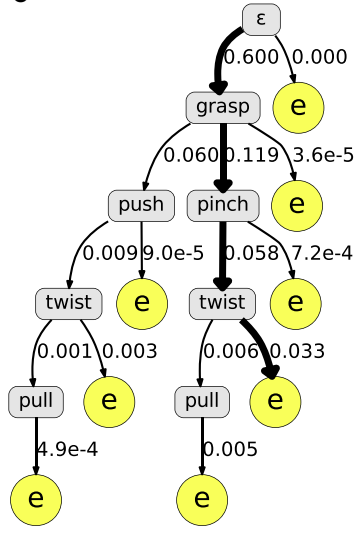



**Fig. 8. An example of the GEP.** (**A**) A classifier is applied to a six-frame signal and outputs a probability matrix as the input. (**B**) Table of the cached probabilities of the algorithm. For all expanded action sequences, it records the parsing probabilities at each time step and prefix probabilities. (**C**) Grammar prefix tree with the classifier likelihood. The GEP expands a grammar prefix tree and searches in this tree. It finds the best action sequence when it hits the parsing terminal *e*. It finally outputs the best label "grasp, pinch, pull" with a probability of 0.033. The probabilities of children nodes do not sum to 1 because grammatically incorrect nodes are eliminated from the search and the probabilities are not renormalized (*22*).

from the analysis for failing to understand the haptic display panel by not passing a recognition task. Hence, the analysis included 150 participants (mean age of 20.7). The symbolic and haptic explanation panels were generated as described in the "Explanation generation" section. The text explanation was generated by the authors based on the robot's action plan to provide an alternate text summary of robot behavior. Although such text descriptions were not directly yielded by the model, they could be generated by modern natural language generation methods.

The human experiment included two phases: familiarization and prediction. In the familiarization phase, participants viewed two videos showing a robot interacting with a medicine bottle, with one successful attempt of opening the bottle and a failure attempt without opening the bottle. In addition to the RGB videos showing the robot's executions, different groups viewed the different forms of explanation panels. At the end of familiarization, participants were asked to assess how well they trusted/believed that the robot had the ability to open the medicine bottle (see text S2.5 and fig. S7 for the illustration of the trust rating question).

Next, the prediction phase presented all groups with only RGB videos of a successful robot execution; no group had access to any explanatory panels. Specifically, participants viewed videos segmented by the robot's actions; for segment *i*, videos start from the beginning of the robot execution up to the *i*th action. For each segment, participants were asked to predict what action the robot would execute next (see text S2.5 and fig. S8 for an illustration of the action prediction question).

Regardless of group assignment, all RGB videos were the same across all groups; i.e., we showed the same RGB video for all groups with varying explanation panels. This experimental design isolates potential effects of execution variations in different robot execution models presented in the "Robot learning" section; we only sought to evaluate how well explanation panels foster qualitative trust, enhance prediction accuracy, and keep robot execution performance constant across groups to remove potential confounding.

For both qualitative trust and prediction accuracy, the null hypothesis is that the explanation panels foster equivalent levels of trust and yield the same prediction accuracy across different groups, and therefore, no difference in trust or prediction accuracy would be observed. The test is a two-tailed independent samples $t$ test to compare performance from two groups of participants, because we used between-subjects design in the study, with a commonly used significance level $\alpha = 0.05$, assuming t-distribution, and the rejection region is $P < 0.05$.

## SUPPLEMENTARY MATERIALS

robotics.sciencemag.org/cgi/content/full/4/37/eaay4663/DC1
Text S1. Additional model results
Text S2. Additional materials and methods
Text S2.1. Model limitations
Text S2.2. Training details of embodied haptic model
Text S2.3. Details on tactile glove
Text S2.4. Force visualization
Text S2.5. Additional human experiment details
Fig. S1. Additional generalization experiments on bottles augmented with different 3D-printed caps.
Fig. S2. Examples of estimated trends for testing and generalization haptic data.
Fig. S3. The confusion matrix of Δ across different bottles based on the haptic signals.
Fig. S4. An example of action grammars and grammar prefix trees used for parsing.
Fig. S5. An example of the GEP.
Fig. S6. Tactile glove hardware design.
Fig. S7. Qualitative trust question asked to human participants after observing two demonstrations of robot execution.
Fig. S8. Prediction accuracy question asked to human participants after each segment of the robot's action sequence during the prediction phase of the experiment.
Table S1. Numerical results and SDs for human participant study.
Table S2. Network architecture and parameters of the autoencoder.
Table S3. Network architecture and parameters for robot to human embedding.
Table S4. Network architecture and parameters for action prediction.
Table S5. Hyperparameters used during training.
Table S6. Specifications of the computing platform used in the experiments.
Algorithm S1. Algorithm of the improved GEP for robot planning.
Movie S1. Example explanation video of full model shown to human participants.
Movie S2. Example haptic explanation video shown to human participants.
Movie S3. Example symbolic explanation video shown to human participants.
Movie S4. Example text explanation video shown to human participants.
Movie S5. Example baseline video shown to human participants.
Data S1. Final data files (.zip).
Code S1. Software files (.zip).

## REFERENCES AND NOTES


1. A. Falcon, Aristotle on causality, in *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition)*, E. N. Zalta, Ed. (Metaphysics Research Lab, Stanford University, 2019).
2. D. Kelemen, The scope of teleological thinking in preschool children. *Cognition* **70**, 241–272 (1999).
3. A. Gopnik, A. N. Meltzoff, P. K. Kuhl, *The Scientist in the Crib: Minds, Brains, and How Children Learn* (William Morrow & Co, 1999).
4. T. Lombrozo, The structure and function of explanations. *Trends Cogn. Sci.* **10**, 464–470 (2006).
5. M. Tomasello, *Origins of Human Communication* (MIT Press, 2010).
6. D. Gunning, *Explainable Artificial Intelligence (XAI)* (Defense Advanced Research Projects Agency, 2017).
7. L. Feng, L. Humphrey, I. Lee, U. Topcu, Human-interpretable diagnostic information for robotic planning systems, paper presented at the International Conference on Intelligent Robots and Systems, Daejon, Korea, 9 to 14 October 2016.
8. B. Hayes, J. A. Shah, Improving robot controller transparency through autonomous policy explanation, paper presented at 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI'17), Vienna, Austria, 6 to 9 March 2017.
9. K. Erol, J. Hendler, D. S. Nau, Complexity results for HTN planning. *Ann. Math. Artif. Intell.* **18**, 69–93 (1996).
10. R. E. Fikes, N. J. Nilsson, Strips: A new approach to the application of theorem proving to problem solving. *Artif. Intell.* **2**, 189–208 (1971).
11. G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
12. Y. Duan, X. Chen, R. Houthooft, J. Schulman, P. Abbeel, Benchmarking deep reinforcement learning for continuous control, paper presented at International Conference on Machine Learning (ICML 2016), New York, NY, 19 to 24 June 2016.
13. I. Lenz, H. Lee, A. Saxena, Deep learning for detecting robotic grasps. *Int. J. Robot. Res.* **34**, 705–724 (2015).
14. G. Marcus, Deep learning: A critical appraisal, http://arxiv.org/abs/1801.00631 (2018).
15. M. Minsky, S. A. Papert, *Perceptrons: An Introduction to Computational Geometry* (MIT Press, 2017).
16. P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (Basic Books, 2015).
17. J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual explanations for self-driving vehicles, paper presented at European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8 to 14 September 2018.
18. C.-K. Yeh, J. Kim, I. E.-H. Yen, P. K. Ravikumar, Representer point selection for explaining deep neural networks, paper presented at 31st Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, 3 to 8 December 2018.
19. R. Hu, J. Andreas, T. Darrell, K. Saenko, Explainable neural computation via stack neural module networks, paper presented at European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8 to 14 September 2018.
20. Q. Zhang, Y. N. Wu, S.-C. Zhu, Interpretable convolutional neural networks, paper presented at Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, 18 to 23 June 2018.
21. K. Tu, M. Pavlovskaia, S.-C. Zhu, Unsupervised structure learning of stochastic and-or grammars, paper presented at 26th Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, 5 to 10 December 2013.
22. S. Qi, B. Jia, S.-C. Zhu, Generalized earley parser: Bridging symbolic grammars and sequence data for future prediction, paper presented at International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 10 to 15 July 2018.
23. M. Edmonds, F. Gao, X. Xie, H. Liu, S. Qi, Y. Zhu, B. Rothrock, S.-C. Zhu, Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open

medicine bottles, paper presented at 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada, 24 to 28 September 2017.
24. T. Lombrozo, Explanation and abductive inference, in *The Oxford Handbook of Thinking and Reasoning* (Oxford Univ. Press, 2013).
25. J. A. Simpson, Psychological foundations of trust. *Curr. Dir. Psychol. Sci.* **16**, 264–268 (2007).
26. J. D. Lewis, A. Weigert, Trust as a social reality. *Soc. Forces* **63**, 967–985 (1985).
27. C. Castelfranchi, R. Falcone, Principles of trust for MAS: Cognitive anatomy, social importance, and quantification, paper presented at Third International Conference on Multi-Agent Systems, Paris, France, 3 to 7 July 1998.
28. H. Liu, X. Xie, M. Millar, M. Edmonds, F. Gao, Y. Zhu, V. J. Santos, B. Rothrock, S.-C. Zhu, A glove-based system for studying hand-object manipulation via joint pose and force sensing, paper presented at International Conference on Intelligent Robots and Systems (IROS 2017), Vancouver, Canada, 24 to 28 September 2017.
29. S.-C. Zhu, D. Mumford, A stochastic grammar of images. *Found. Trends Comput. Graph. Vis.* **2**, 259–362 (2007).
30. J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergonomics* **4**, 53–71 (2000).
31. J. Earley, An efficient context-free parsing algorithm. *Commun. ACM* **13**, 94–102 (1970).
32. A. Stolcke, An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *J. Comput. Linguist.* **21**, 165–201 (1995).

**Acknowledgment:** We are grateful to the editor of the special issue and the two reviewers for their valuable comments that have helped improve the presentation of the paper. We thank W. B. Wu (Department of Statistics, University of Chicago) for insightful discussions on data analysis and Q. Xie for running the human experiment. We also thank B. Jia, Z. Zhang, and C. Zhang (UCLA Computer Science Department) for helpful discussions. **Funding:** This work was supported by DARPA XAI N66001-17-2-4029. **Author contributions**: M.E.: data collection, designing and running participant study, writing, and building robot platform; F.G.: student lead of running robot experiments, model formulation and coding, cleaning tactile data, running participant study, and building robot platform; H.L.: building tactile glove and coding, data collection, running participant study, and writing; X.X.: cleaning tactile data, running participant study, and building tactile glove and coding. S.Q.: model formulation, running the human experiment, and writing. B.R.: building the robot platform and coding, and writing. Y.Z.: building the robot platform and coding, and writing. Y.N.W.: model formulation and writing. H.L.: designing and running the human experiment and writing. S.-C.Z.: investigating the key idea and providing the environment and the funding support for conducting this research. **Competing interests:** M.E., X.X., Y.Z., and S.-C.Z. have affiliations with the International Center of AI and Autonomy (CARA); H.L. and S.-C.Z. have affiliations with DMAI Inc.; S.Q. is currently an employee of Google LLC; B.R. is currently an employee of Paige.AI 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 entirely conducted at UCLA and JPL. **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.

Submitted 21 June 2019
Accepted 26 November 2019
Published 18 December 2019
10.1126/scirobotics.aay4663

**Citation:** 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**, eaay4663 (2019).

# A tale of two explanations: Enhancing human trust by explaining robot behavior

Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu and Song-Chun Zhu

*Sci. Robotics* **4**, eaay4663.
DOI: 10.1126/scirobotics.aay4663