Influencing Behavioral Attributions to Robot Motion During Task Execution

Anonymous Author(s)
Affiliation
Address
email

Abstract: Recent literature has proposed algorithms for autonomous generation of robot motion that communicates functional attributes of a robot's state such as intent or incapability. However, less is known about how to automate the generation of motion for communicating higher-level behavioral attributes such as curiosity or competence. We propose a framework that jointly addresses the challenges of modeling human attributions to robot motion, generating trajectories that elicit attributions, and selecting trajectories that balance attribution and task completion. The insight underpinning our approach is that many attributions can be traced to salient features of the robot's motion. We illustrate the framework in a coverage task resembling household vacuum cleaning. Through a virtual interface, we collect a dataset of human attributions to robot trajectories during task execution and learn a probabilistic model that maps robot trajectories to human attributions. We then incorporate this model into a trajectory generation mechanism that balances between task completion and communication of a desired behavioral attribution. Through an online user study on a different household layout, we find that our prediction model accurately captures human attribution for coverage tasks.

1 Introduction

As robots enter households and public spaces, it is increasingly important to account for human perceptions of their behavior [1, 2, 3, 4]. While a robot's actions might be driven by unambiguous internal objectives, solely optimizing such criteria might result in robot behavior that is difficult to interpret or disruptive to human onlookers. For example, a highly articulated robot may follow a non-humanlike trajectory that users attribute caprice to, making observers uncomfortable [5] or a home robot such as a robot vacuum cleaner can exhibit unpredictable motion that causes observers to perceive the robot as broken, interrupting home activity.

Accounting for high-level attributions to robot behavior is a complex problem relying on the mechanisms underlying human inference and behavior generation. In psychology, there is a long history of work on understanding human attribution for behavior explanation or inference of behavior traits [6, 7]. Humans are highly attuned to how their actions are perceived and adapt their behavior to elicit a desired impression from others or adhere to social norms, a concept that is known as *presentation of self* [8]. The tendency for humans to attribute even situational behaviors to deeper character traits is so pervasive that it is known as "the fundamental attribution error" [9].

Inspired by these theories, we envision robots that can 1) leverage an understanding of humans' attribution mechanisms to predict attributions to their motion, 2) generate behaviors that elicit desired human impressions and 3) balance attribution elicitation and task completion. This will increase the acceptance of robots in human spaces by enabling them go about their tasking in a way that is both effective and sensitive to perceptions.

We propose a framework that addresses these challenges by extracting the salient features of a robot's motion and the attributions they elicit from users and integrating them into the a generation scheme that enables the flexible tradeoff of task completion and attribution. We illustrate this framework in a virtual representation of a robot vacuum cleaning task. We learn a probabilistic

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

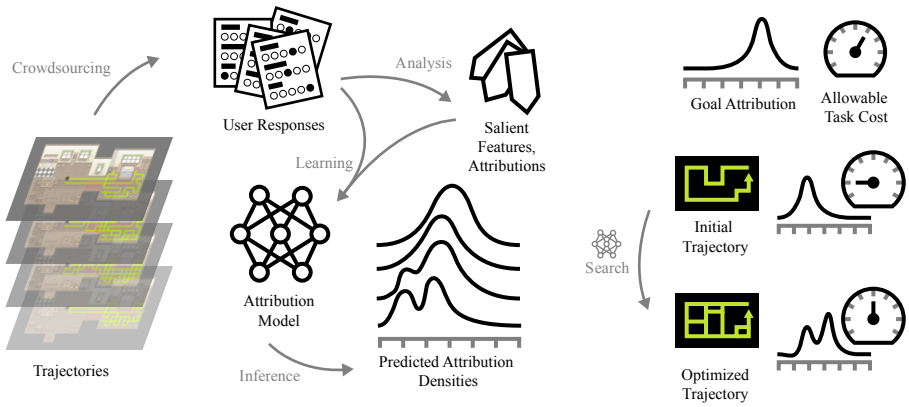


Figure 1: Our proposed framework. User responses to robot trajectories are analyzed to extract salient features and attributions and used to train a model that probabilistically maps robot trajectories to human attributions (left). The acquired model is used to generate robot trajectories that elicit a desired attribution (right).

model that maps a robot coverage trajectory to the expected distribution of human ratings. We then incorporate this model into a trajectory generation scheme that balances task-cost and attribution elicitation objectives. We generate a range of trajectories optimizing for different attributions with varying cost constraints, demonstrating that the approach produces thematically consistent trajectories in both an environment it was trained in and an environment variant. We conduct an online user study and verify that attributions to the generated trajectories are well predicted by the model.

2 A Framework for Behavioral Attribution

We consider a robot performing a task $\mathcal{T}$ in a human environment. We denote by $s \in \mathcal{S}$ the robot state where $\mathcal{S}$ is a state space and define a robot trajectory as a sequence of states $\xi = (s_0, \ldots, s_t)$ where indices correspond to timesteps following a fixed time parametrization of step size. Let us define the task as a tuple $\mathcal{T} = (\Xi, \mathcal{A}, \mathcal{P}, C)$ where Ξ is a space of robot trajectories, $\mathcal{A}$ denotes the robot action space, $\mathcal{P} : \Xi \times \mathcal{A} \to \Xi$ represents a deterministic state transition model, and $\mathcal{C} : \Xi \to \mathbb{R}$ is a trajectory cost.

We assume that the robot starts from an initial state s_0 and reaches a terminal state s_T at the end of the task execution (time T) by executing a trajectory $\xi = (s_0, \ldots, s_T)$. We assume that this trajectory ξ is fully observable by a human observer who is aware of the task specification $\mathcal{T}$. The observer makes an inference of the form $\mathcal{I}_B : \Xi \times \mathcal{T} \to \mathcal{B}$, where Ξ is a space of trajectories and $\mathcal{B}$ is a space of behavioral attributions. The form of $\mathcal{B}$ will vary, but should be selected to capture the range, combinations, and intensities of attributions that the robot should be sensitive to.

Conversely, given a behavioral attribution $\boldsymbol{b}$ from a space of behavioral attributions $\mathcal{B}$ and a task $\mathcal{T}$, the observer expects to see a trajectory $\xi_b \in \Xi$, corresponding to an inference of the form $\mathcal{I}_\xi : \mathcal{B} \times \mathcal{T} \to \Xi$. In other words, we assume that there is a "way" that a curious—or any other attribution—robot would be expected to execute a particular task. Equivalently, we assume that there is a set of trajectories that the observer could more often imagine to be consistent with some attribution.

In the remainder of paper, we aim to provide a general framework for modeling inferences of the form $\mathcal{I}_B$, and $\mathcal{I}_\xi$. Our goal is to enable robots to understand and account for the communicative effects of their motion on human observers.

3 Related Work

Several studies have illustrated the value of robot motion as a communicative modality [10, 11, 12, 13, 14, 15]. Some works propose algorithms for legible robot motion generation, which have been shown to enable effective human-robot collaboration in manipulation tasks [10], or smooth robot navigation in close proximity to humans [12, 14]. Other works focus on conveying higher-level

information such as the robot's objective function [15] or the source of failure [13] in failure cases. Animation principles [16] or movement analysis [17] are often employed to inform the design of expressive robot behaviors. Finally, related graphics research focuses on the generation of stylistically distinct but functionally equivalent motion primitives for walking and other activities [18, 19].

The complex interplay of embodiment and communicative motion has motivated research on understanding human perceptions of robot behavior. For instance, early work looked at the effect of robot gaze on human impressions [20]. Sung et al. [1] study human attitudes towards robot vacuum cleaners and propose design principles aimed at enhancing the acceptance of robots in domestic environments. [2] report a relation between robot motion and perceived affect. Lo et al. [21] and [14] investigate human perceptions of different robot navigation strategies whereas Walker et al. [22] study human perceptions of robot actions that deviate from the robot's assigned task.

Our work draws inspiration from recent work on design and methods for extracting human perceptions and attributions for robot motion [1, 2]. However, it moves beyond the problem of understanding and analyzing human perceptions, and focuses on the problem of *synthesizing* implicitly communicative motion. Our work is closely related to past work on the generation of legible robot motion [10, 11, 23, 13] in that we also incorporate a model of human inference into the robot's motion generation pipeline. However, unlike these works which emphasize the communication of task-related attributes, our focus is instead on communicating high-level, behavioral attributes through robot motion.

4 Modeling and Influencing Attributions

We consider a scenario in which a mobile robot performs a coverage task in a two-dimensional discrete workspace while a human is observing from a top-down view. We employ a virtual environment[1] that resembles a house and stylize the agent as a robot vacuum cleaner (see Fig. 2) since the general population is already somewhat familiar[2] with such robots [1, 24], making it easier for participants to develop mental models about their motion than that of a manipulator, for example.

In this scenario, the robot state space is the complete home workspace and Ξ is the space of all possible trajectories of any length that could be followed in the space. The robot action space $\mathcal{A}$ consists of the cardinal directions and actions are deterministic. The cost of a state transition from a state s_t to a state s_{t+1} after having followed a trajectory ξ_t is defined as 0 if the state hasn't been visited before, -5 if the state contains a small, traversable obstacle (e.g. a vase), and -1 otherwise. A penalty of 3 times the number of unvisited states from the goal region is applied on termination.

4.1 Understanding Behavioral Attribution for Coverage Tasks

Through exploratory studies on Amazon Mechanical Turk, we sought to extract domain knowledge for attributions to robot motion within coverage tasks. Using the home layout of Fig. 2, we generated a set of trajectories exhibiting qualitatively distinct ways the robot could respond to the prompt to "clean the bedroom," ranging from a near optimal coverage plan to a trajectory that barely visited the target room (see Appendix A for additional details). Each participant viewed videos of a random selection of three of these trajectories. After each video, participants were asked: a) to provide three words to describe the robot's behavior; b) to rate their agreement that "the robot is _____" for a range of adjectives drawn from relevant literature on human attributions [25, 26, 22]; c) to "explain what factors contributed to their strongest ratings." In addition to attributions, participants were asked to use an interactive interface to demonstrate how they would "clean the bedroom in a way that makes the robot look _____" where the blank was filled at random with an adjective from the attribution rating items. Across all exploratory studies, we collected 375 sets of attribution ratings from 115 participants (73 male, 41 female) aged 21-70 (M = 38.3, SD = 10.7) covering 63 trajectories and a total of 193 demonstrations.

[1] The environment is built in the Phaser game engine (https://phaser.io/) and uses art by Bonsaiheldin under a CC-BY-SA license.

[2] A presentation by iRobot [24] reports that about 14M households had robot vacuum cleaners in the U.S. in 2019.

Feature	Description
Coverage (%)	Goal region states visited at least once.
Redundant coverage (%)	Goal region states visited more than once.
Overlap (%)	Plan states visited more than once.
Length (%)	Normalized plan length.
Hook template (%)	Frequency of "U" shape patterns in plan.
Straight template (%)	Frequency of action repetition in plan.
Start-stop template (%)	Frequency of idle-move-idle patterns in plan.
Idleness (%)	Frequency of idle actions in plan.
Map coverage (%)	Ratio of map states visited at least once.
Collision (%)	Ratio of obstacle states from $\mathcal{O}$ in plan.
Goal deviation (%)	Ratio of of plan before first goal state.

Table 1: Low-dimensional trajectory representation.

Figure 2: The home environment used in our exploratory studies.

4.1.1 Extracting the Space of Attributions

To understand the inter-correlation of participant adjective ratings, we conducted an exploratory factor analysis. We selected a three-factor, promax rotation model which explained 74% of the observed variance due to its parsimony and coherence (see Appendix B). The first factor, which we call "competence" for its similarity to the relevant factor described by Carpinella et al. [26] consists of six items (responsible, competent, efficient, reliable, intelligent, focused) centered on the capability and diligence of the robot. The second consists of four items (lost, clumsy, confused, broken) alluding to a temporary or extended negative state, for which we title the factor "brokeness". The third contains two items (curious, investigative) and matches the curiosity factor examined by Walker et al. [22].

The extracted model enables the computation of standardized factor scores roughly in the range $[-3, 3]$ which denote how many standard deviations from the mean a participant's ratings for the items are. Reflecting the format of the component items, a high or low factor score denotes agreement or disagreement that a trajectory expresses an attribution, respectively. Based on this model, we represent the attribution for a trajectory ξ as a tuple $\boldsymbol{b} = (b_{\textit{competent}}, b_{\textit{broken}}, b_{\textit{curious}}) \in \mathcal{B}$ where the space of attributions is the set $\mathcal{B} = [-3, 3]^3$.

4.1.2 Low-dimensional Trajectory Representation

The space of possible trajectories in this domain is too large to explore directly, so we constructed a low-dimensional space Φ based on features relevant to the formation of attribution ratings. This allows us to describe a trajectory ξ as a vector $\phi_\xi = \phi(\xi) \in \Phi$. The feature space was inspired by relevant literature on human behavioral attribution to robot motion and enriched with features appearing in participants' explanations. The final set of 11 features used in further experiments is listed in Table 1.

4.2 Mapping Trajectories to Attribution Scores

Given a trajectory ξ, an observer's inference $\mathcal{I}_B$ of behavioral attribution can be expected to vary both due to individual differences and as a result of measurement error. For this reason, we model $\mathcal{I}_B$ as a conditional probability density $f_{\mathcal{B}|\Xi}(\boldsymbol{b}|\phi_\xi) : \mathcal{B} \to \mathbb{R}$. We observed multimodality in the distribution of factor scores for some trajectories, so we use a Mixture Density Network (MDN) [27] to approximate each conditional density as a mixture distribution $f_{\mathcal{B}|\Xi}(\boldsymbol{b}|\phi_\xi) = \sum_{i=1}^{C} \alpha_i(\phi_\xi) k_i(\boldsymbol{b}|\phi_\xi)$ where α_i, $i = 1, \ldots, C$, is a mixing coefficient, and k_i is a multivariate Gaussian kernel function with mean $\boldsymbol{\mu}_i$ and covariance Σ_i. Note that the mixing coefficients α_i and the Gaussian parameters $\boldsymbol{\mu}_i$ and Σ_i are functions of the featurized trajectory ϕ_ξ. In our models, these functions are implemented as linear transformations of features produced by a shared multi-layer perceptron.

To make efficient use of scarce data, we created ensembles of MDNs using bootstrap aggregation, i.e., we trained N models with different data splits and uniformly weight their predictions: $f^{ens}_{\mathcal{B}|\Xi}(\boldsymbol{b}|\phi_\xi) = \frac{1}{N} \sum_{i=1}^{N} f^{i}_{\mathcal{B}|\Xi}(\boldsymbol{b}|\phi_\xi)$.

We studied three different model configurations; single and four component MDNs, i.e., $C = 1$ and $C = 4$ and an ensemble of 8 MDNs each with four components, i.e., $C = 4$, $N = 8$. We trained all models using an average negative log likelihood (NLL) loss function, the Adam optimizer [28],

Model	Parameters	Average NLL	SD
Uniform	6	5.38	0.00
MDN, C=1	120	$3.13 \pm .05$	$1.35 \pm .09$
MDN, C=4	300	$2.66 \pm .08$	$1.57 \pm .05$
MDN Ensemble, C=4 N=8	2400	$2.53 \pm .06$	$1.38 \pm .04$

Table 2: Average test negative log likelihood (NLL) for each model configuration. Each datapoint represents a mean NLL over 16 models trained with random train-validate folds on a fixed test set. Error is the 95% confidence interval calculated with bootstrapping.

noise regularization [29], and early stopping. We configured the input MLP to use a single hidden layer with 5 units and a hyperbolic tangent activation. We expanded the dataset collected in our exploratory studies after assessing the sensitivity of the models to increased amounts of data, a process described in Appendix C. The final version of the set includes 126 trajectories with 671 attribution ratings. Table 2 compares the NLL of the models over held-out data. The mean indicates the typical quality of the prediction and the standard deviation indicates the degree to which this varied from sample to sample. Both quantities are averaged over 16 random folds and reported with bootstrapped 95% confidence intervals. All models perform significantly better than a uniform baseline, which simply assigns equal probability to all outcomes. The ensemble model performs best and is used in further experiments in the remainder of the paper.

4.3 Generating Trajectories that Elicit Desired Attributions

We represent the behavior specification as a one-dimensional Gaussian $b^* \sim \mathcal{N}(\mu_b, \sigma_b^2)$ centered on a desired rating $\mu_b \in [-3, 3]$ for a single attribution dimension where the variance σ_b serves as a tolerance parameter modeling the acceptable distance from the desired behavioral rating. We use a density representation as it more closely matches the output of our model for $\mathcal{I}_B$.

Together with the task requirements as described by the cost function $\mathcal{C}$, we realize the inference $\mathcal{I}_\xi$ as an optimization scheme of the form:

$$\begin{aligned} \xi^* = \arg\min_{\xi \in \Xi} \quad & D_{\mathrm{KL}}(f_{\boldsymbol{B}_i} || \mathcal{N}(\mu_b, \sigma_b^2)) \\ \text{s.t.} \quad & \mathcal{C}(\xi) \leq w, \end{aligned} \tag{1}$$

where D_{KL} denotes the KL divergence, $f_{\boldsymbol{B}_i}$ is the density $f_{\boldsymbol{B}|\Xi}|\phi(\xi)$ marginalized across dimensions other than i, and w is the maximum allowable task cost. Because many applications are conventionally exclusively task-cost driven, this format provides an intuitive "knob" in the form of how suboptimal the robot is allowed to be. Where performance is critical—perhaps to meet a schedule or to fit in power constraints—the robot designer need only set w to match the need. Different attributions are expected to be more or less sensitive to the allowable suboptimality, something illustrated in Appendix D.

We implement this optimization using a hill-descending search in the space of trajectories. The search is initialized with a task-optimal trajectory generated via A* search, and progressively samples modifications to the trajectory. These modifications consist of both naive, action-level modifications to the trajectory as well as changes targeting the activation of the features underlying the attribution model (see Fig. 1). Important motion templates, like runs of straight motion, hook patterns, start-stops are sampled and patched into trajectories. All modifications are ranked by the divergence of their predicted attribution with the behavior specification. The search terminates after a fixed duration and the best performing trajectory subject to the task-cost constraint is returned. A detailed description of the optimization procedure is given in Appendix D.

5 Evaluation

We conduct a user study to evaluate the efficacy of the framework as a means of producing trajectories that elicit desired attributions. Our study is motivated by the following hypotheses:

Hypothesis 1 The model makes accurate predictions about the distribution of attributions to new trajectories.

Hypothesis 2 The model makes accurate predictions about the distribution of attributions to trajectories in unseen environments.

Hypothesis 3 The approach enables the generation of trajectories that elicit desired attributions.

5.1 Experiment Design

In Experiment I, participants observe and rate trajectories in the same home layout used for data collection, while in Experiment II, trajectories are generated in a modified home layout. In an effort to understand the impact of the environment geometry, the modified layout increases the size of the goal region by 100%, varies the placement of items and obstacles, and flips the dominant direction of the robot's motion. The experiments are within-subjects, video-based user studies, both instantiated in three parallel sets corresponding to the three attribution dimensions considered. For each dimension, we consider four robot trajectories generated by optimizing (1) with a target distribution expressed as a Gaussian centered at 1.5 with scale 0.3 and varying task cost thresholds. To ease interpretation, the w values governing the thresholds were set in multiples—**1x**, **2x**, **4x**, **12x**—of the cost of the optimal trajectory for the task. The full set of trajectories is shown in Fig. 3. In all experiments, participants rate and describe each trajectory using the same items and questions used in the exploratory studies of Sec. 4. After watching all trajectories in a randomly assigned order, they also respond to additional comparative questions: "which robot seemed the most _____" and "which robot seemed the least _____", where the blanks are filled with the adjective corresponding to the dimension of attribution studied. Both comparisons are accompanied with an open-ended question asking for a brief explanation of the choice. No suitable baselines exist for the balanced attribution elicitation task, so our experiments use solely trajectories generated by our approach.

Participants A total of 144 participants (76 male, 68 female) aged 20-72 (M = 38.2, SD = 10.9) were recruited via Amazon Mechanical Turk and paid $2 to complete the approximately 15 minute task. 9 had taken part in our earlier exploratory studies. Participants were equally distributed amongst the six sets of conditions. Condition orderings were fully counterbalanced.

5.2 Results

The predictive performance of the model is illustrated in Fig. 5 and reported in Table 3. The 95% bootstrapped confidence interval on the mean NLL for Experiment I was 2.77$\pm$.10 (SD = 0.88$\pm$.09). and 2.87$\pm$0.10 (SD = 0.86$\pm$.09) for Experiment II. Participants' choices for "most" and "least" trajectories are shown in Fig. 4.

Hypothesis 1 was supported; the average NLL of the models was significantly lower than a uniform model, indicating that the model was able to meaningfully predict attributions in the layout it was trained in.

Hypothesis 2 was supported; the average NLL of the attributions observed across Experiment II was significantly lower than the uniform model, indicating that the model remains informative even under modifications to the task layout.[3]

Hypothesis 3 saw mixed support; Kendall's tau-b correlation tests (reported in Appendix E) indicate strong positive correlations between the allowable suboptimality and brokenness factor scores, suggesting that the trajectory generation method was effective at eliciting progressively higher factor scores. However, tests for the competence conditions indicated a moderate negative correlation, and tests for curiosity conditions were not significant. As shown in Fig. 4, while participants found that 12x and 1x were the most and least "_____" for experiments focused on curiosity and brokenness, this relationship was flipped for the competence experiments.

When optimizing for competence, the model emphasizes over-coverage of the goal region as well as coverage of the house as a whole (see Fig. 3). Due to the associated task-cost penalty, coverage outside of the goal begins to appear in the 4x condition of Experiment I and the 12x condition of Experiment II, and participants' responses indicate that it is a key driver of negative attributions of competence. Some emphasized that time spent not cleaning the bedroom was "wasted movement

[3] See Appendix E for supplementary tests indicating insufficient evidence to support a significant difference between the predicted and observed distributions in both experiments.

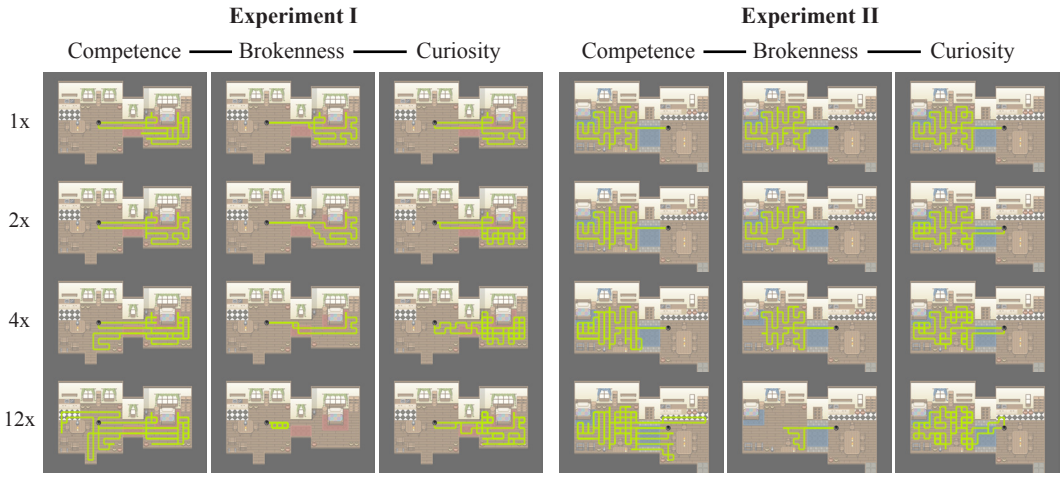


Figure 3: Traces of robot trajectories used in different experiments and conditions.

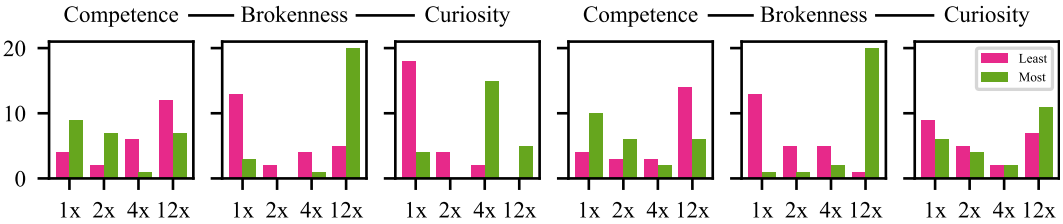


Figure 4: Counts of trajectories picked as most and least "_____".

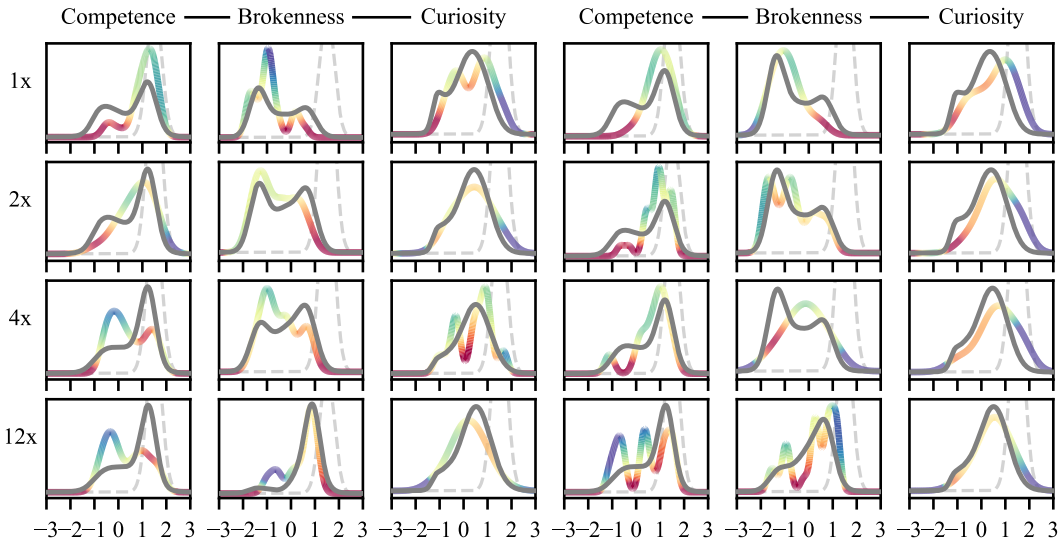


Figure 5: Comparisons of the model's predicted density for the factor score under consideration (in grey), the observed distribution (multicolored) approximated with Gaussian kernel density estimation and the target distribution (dashed). Each subplot has a unique y-scale, with the magnitude of the difference between the predicted and observed densities at any point encoded instead in the color of the line for the observed distribution. Deep red indicates the model severely overpredicted the density, while deep blue indicates severe underprediction.

	Competence	Brokenness	Curiosity	Competence	Brokenness	Curiosity
1x	2.51 (0.69)	2.44 (0.76)	2.37 (0.54)	2.73 (0.59)	2.64 (0.73)	2.76 (1.05)
2x	2.76 (0.85)	2.84 (0.72)	2.85 (0.75)	2.95 (0.70)	2.86 (0.88)	2.85 (1.09)
4x	2.98 (0.76)	2.93 (0.83)	3.02 (0.71)	3.00 (0.67)	2.92 (0.42)	3.10 (0.98)
12x	3.10 (0.73)	2.53 (1.75)	2.94 (0.72)	3.43 (0.84)	2.28 (0.98)	2.84 (0.84)

Table 3: Evaluation Average NLL (SD)

in the wrong room" (Exp. I-Competence), while others attributed the deviation to being "lost" or "totally confused" (Exp. I-Competence). In less extreme conditions, users expressed uncertainty about what drove the behavior, saying they were "not sure if robot is cleaning outside the bedroom cause there might be dirt that can be brought in, or just confused as to parameter of bedroom" (Exp. I-Competence, 4x). A minority of users thought that the extra motion was worthwhile, marking the 12x trajectory as the most competent because it "cleaned more areas in both rooms" or "completed the entire home from bedroom to kitchen" (Exp. I-Competence). The model overestimated the prevalence of people that would appreciate the additional coverage of the environment, as indicated by both the "most/least" selections and the NLL. We speculate that the format of the experiment—wherein participants view the optimal trajectory and mentally anchor their ratings against it—is a contributing factor; participants may be more inclined to rate the robot as competent when viewing the optimized trajectories in real-world settings.

Trajectories optimized to look broken progressively cover less of the goal region before ultimately devolving into repeated circular motion near the start point (see Fig. 3). The majority of users concurred in their assessment of the 12x condition as "defective", with some remarking that it seemed overwrought, "a joke version of the robot" (Exp. II-Brokenness). The model underpredicted the extent of participants' agreement that the optimal trajectory would be perceived as not broken, but the predictive performance across all brokenness conditions was still the strongest of the three factors.

When optimizing for curiosity, the model emphasized over-coverage of the goal region, overlapping motion, hook-like patterns and visiting penalized states depicted with vases (see Fig. 3). Some participants highlight the extra coverage as the reason for selecting the 12x as the most curious condition, saying the robot "cleans very well but cleaned the same place multiple time, roaming without reason" or that "dawdled around a lot, getting hyper fixated on certain spots" (Exp. II-Curiosity). The same factor is highlighted by participants in less extreme conditions, with one speculating of the 4x trajectory that "maybe something caught its eye while it was working and it got so distracted that he totally kept getting off track" (Exp. I-Curiosity). The change in the distribution of curiosity factor scores was expected to be small and the observations bear the predictions out, though it is notable that despite the subtle differences a majority of participants select either 4x or 12x as the "most curious" trajectory across both experiments.

6 Discussion

Our framework is a general approach for endowing robots with a sensitivity to the behavioral attributions their motion elicits. We illustrated its application to coverage motions, but we envision a lively stable of related instances stretching across domains from delivery service robots to robot arms in fulfillment centers. The process is the same; the robot arm's designer will build a pool of videos and study users' responses to—and their reasoning about—the motion over a broad set of dimensions, then learn a forward mapping from the features driving their reasoning to the attributions in their responses. While some features such as path length or redundancy may map over from the coverage domain, others like the shape of the acceleration profile may need to be added to capture perceptions of danger or erraticism. The reverse translation from a desired attribution to a new trajectory can be realized by searching in the space trajectories, using the attribution features to guide the process—something that may have a pronounced impact in a higher dimensional planning space. This optimization can be naturally constrained in terms of the maximum permissible task-cost of the trajectory.

While our results showed that the framework's performance transferred to a similar environment, exploring its application to alternative settings is important future work. In the setting we explored, the observer looks at the robot's motion from a top-down perspective, but it's often the case that an observer will watch a robot's motion from a different perspective, which may result in different impressions. Towards our goal of employing this framework on a real robot, we will explore data collection techniques and environments that account for variability in observer perspectives. The features and the learned mappings from features to attributions are specific to the types of environments and the task considered and would likely need to be augmented to work more broadly. In the future, we imagine that robot designers will have access to a wide array of well-studied features with which to bootstrap their system, and when human-robot interaction data is abundant, we may see the rise of learned representations that can power the understanding and generation of motion with minimal additional supervision.

References

[1] J.-Y. Sung, L. Guo, R. E. Grinter, and H. I. Christensen. "my roomba is rambo": Intimate home appliances. In J. Krumm, G. D. Abowd, A. Seneviratne, and T. Strang, editors, *UbiComp 2007: Ubiquitous Computing*, pages 145–162. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[2] M. Saerbeck and C. Bartneck. Perception of affect elicited by robot motion. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 53–60, 2010.

[3] M. Kwon, M. F. Jung, and R. A. Knepper. Human expectations of social robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 463–464, 2016.

[4] E. Cha, A. D. Dragan, and S. S. Srinivasa. Perceived robot capability. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 541–548, 2015.

[5] F. Zacharias, C. Schlette, F. Schmidt, C. Borst, J. Rossmann, and G. Hirzinger. Making planned paths look more human-like in humanoid robot manipulation planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1192–1198, 2011.

[6] F. Heider. *The psychology of interpersonal relations*. John Wiley & Sons Inc., 1958.

[7] B. F. Malle. Attribution theories: How people make sense of behavior. In D. Chadee, editor, *Theories in social psychology*, page 72–95. Wiley Blackwell, 2011.

[8] E. Goffman. *The Presentation of Self in Everyday Life*. Anchor, 1959.

[9] L. Ross. The intuitive psychologist and his shortcomings: Distortions in the attribution process. volume 10 of *Advances in Experimental Social Psychology*, pages 173–220. Academic Press, 1977.

[10] A. Dragan and S. Srinivasa. Integrating human observer inferences into robot motion planning. *Autonomous Robots*, 37(4):351–368, 2014.

[11] R. A. Knepper, C. Mavrogiannis, J. Proft, and C. Liang. Implicit communication in a joint action. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 283–292, 2017.

[12] D. Carton, W. Olszowy, and D. Wollherr. Measuring the effectiveness of readability for mobile robot locomotion. *International Journal of Social Robotics*, 8(5):721–741, 2016.

[13] M. Kwon, S. H. Huang, and A. D. Dragan. Expressing robot incapability. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, page 87–95, 2018.

[14] C. Mavrogiannis, A. M. Hutchinson, J. Macdonald, P. Alves-Oliveira, and R. A. Knepper. Effects of distinct robotic navigation strategies on human behavior in a crowded environment. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 421–430, 2019.

[15] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan. Enabling robots to communicate their objectives. *Autonomous Robots*, 43(2):309–326, 2019.

[16] L. Takayama, D. Dooley, and W. Ju. Expressing thought: Improving robot readability with animation principles. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, page 69–76, 2011.

[17] H. Knight and R. Simmons. Expressive motion with x, y and theta: Laban effort features for mobile robots. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 267–273, 2014.

[18] M. Brand and A. Hertzmann. Style machines. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, page 183–192. ACM Press/Addison-Wesley Publishing Co., 2000.

[19] S. J. Guy, S. Kim, M. C. Lin, and D. Manocha. Simulating heterogeneous crowd behaviors using personality trait theory. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 43–52, 2011.

[20] T. Kanda, H. Ishiguro, and T. Ishida. Psychological analysis on human-robot interaction. In *Proceedings IEEE International Conference on Robotics and Automation (ICRA)*, volume 4, pages 4166–4173 vol.4, 2001.

[21] S. Lo, K. Yamane, and K. Sugiyama. Perception of pedestrian avoidance strategies of a self-balancing mobile robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1243–1250, 2019.

[22] N. Walker, K. Weatherwax, J. Alchin, L. Takayama, and M. Cakmak. Human perceptions of a curious robot that performs off-task actions. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, page 529–538, 2020.

[23] C. Mavrogiannis, W. B. Thomason, and R. A. Knepper. Social momentum: A framework for legible navigation in dynamic multi-agent environments. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 361–369, 2018.

[24] iRobot. *Investor Presentation - August 2019*. August 2019. URL https://investor.irobot.com/static-files/cd40f151-aede-4047-a24f-7e14d317723d. Accessed: 2021-03-01.

[25] C. Bartneck, E. Croft, and D. Kulic. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[26] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner. The robotic social attributes scale (rosas): Development and validation. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 254–262, 2017.

[27] C. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, 1994.

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[29] J. Rothfuss, F. Ferreira, S. Boehm, S. Walther, M. Ulrich, T. Asfour, and A. Krause. Noise regularization for conditional density estimation. *arXiv:1907.08982*, 2019.

A Initializing a Pool of Trajectories

Our approach suggests beginning with a pool of trajectories, however obtaining this pool can pose a bootstrapping problem. It's unlikely that a diverse corpus of trajectories—one that would span the feasible attribution space and enable a data-driven extraction approach—would already exist. In order to create such a corpus, one would need to be confident that they have included trajectories that adequately cover the range of attributions that are perceivable in a domain, but the lack of such information is why we want the corpus of trajectories to begin with.

We broke this cycle by referencing existing literature on the perceptions of robot vacuum cleaners to select six adjectives we considered as candidate attributions: curious, broken, energetic, lazy, lost, scared. We then manually demonstrated six trajectories that we judged to maximally express these candidates. We collected responses to these trajectories and conducted an initial analysis of what aspects of the motion users said contributed to their ratings. This informed the creation of an initial subset of the features we would use. Additional trajectories were then generated using a hill-descending search in the space of trajectories optimizing for incrementally altering individual features ($\pm$0.1-0.3) selected at random while holding other features constant. We found that alternative optimization criteria, like diversity, would lead to degenerate or trivial trajectories which happened to have extreme feature values. These new trajectories were posed to users for their ratings. Simultaneously, we asked them to demonstrate how they would "clean the bedroom in a way that makes the robot look _____" for random items from our questionnaire. This feedback fed back into analysis and the process was cycled through two more times with a subset of the generated and demonstrated trajectories, resulting in the final exploratory dataset and trajectory featurization.

B Loadings

The factor loadings of our final model, derived via exploratory factor analysis of the exploratory dataset, are shown in Table 4.

Variable	Factor 1	Factor 2	Factor 3	Communality
Responsible	**1.00**	.06	−.07	1.01
Competent	**.94**	.03	.00	.89
Efficient	**.93**	.05	−.03	.87
Reliable	**.85**	−.05	.07	.73
Intelligent	**.85**	−.05	.06	.73
Focused	**.84**	−.07	.02	.71
Lost	−.03	**.90**	.01	.80
Clumsy	.13	**.76**	.06	.59
Confused	−.21	**.76**	.03	.63
Broken	−.17	**.62**	−.11	.43
Curious	−.04	.06	**.91**	.83
Investigative	.12	−.02	**.78**	.62

Table 4: Factor loading matrix

C Data Sensitivity

We investigated the sensitivity of our candidate models' performance on unseen trajectories to increased amounts of training data. We divided our final dataset, holding out 20% of the trajectories, then trained and tested models using varying percentages of the training data, repeating the evaluation with 8 randomized folds of the training data. The results, shown in Fig. 6, indicate diminishing returns beginning around the use of 50% of the training data.

D Trajectory Optimization

We would like to generate trajectories that elicit a particular attribution according to the optimization (1). We have the forward model $f_{\boldsymbol{B}|\Xi}|\phi(\xi)$ which predicts a distribution over attributions given

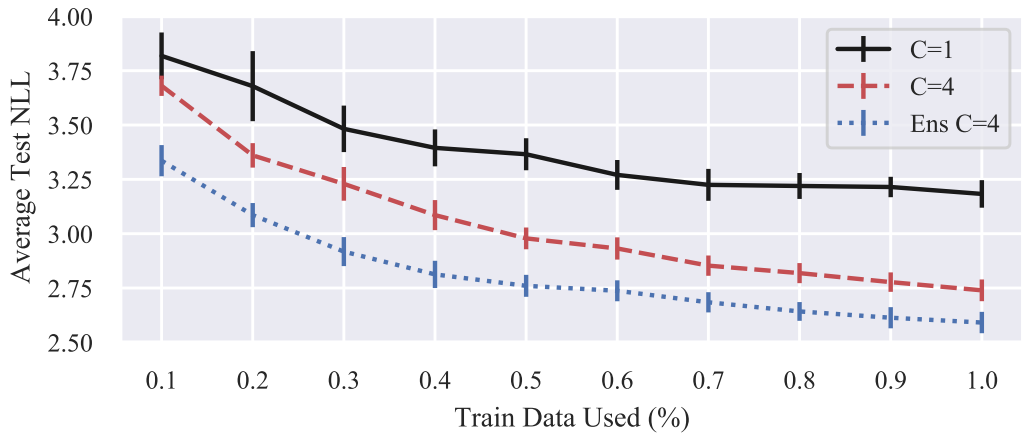


Figure 6: Average test negative log likelihood as a function of the amount of labeled human data used in training. Each datapoint represents the mean of the average test NLL over 10 random folds. Error bars denote standard deviation.

the featurization of a particular trajectory. This model affords two routes towards realizing the optimization. In the first, we directly optimize the features to maximize the objective. However, in that case, a further optimization would be needed to find a trajectory that has the desired feature vector, and it is possible (and common in practice) that there does not exist a trajectory that produces the target features. We instead adopt an alternative approach, in which we search in the space of trajectories. This space is large, and in general there is no clear best way to structure the search. Fortunately the features we use have a clear relationship with patterns of motion, so we can easily sample trajectories that increase the activation of individual features. Given a trajectory, we sample a large set of neighboring trajectories using the following modifications:

Action modification : The trajectory is scanned and new trajectories are initialized by individually witholding each state s_i. The single removed state is replaced by selecting every valid alternative action from s_{i-1} and reconnecting the trajectory with a shortest path to s_{i+1}.

Shortcutting : Sections of the trajectory are deleted uniformly at random and the trajectory is reconnected with a shortest path. We sample twice as many cuts as there are states in the trajectory. Shortcutting is a well established post-processing step in motion planning, helping to uncover shorter or cheaper trajectories that still satisfy the objective.

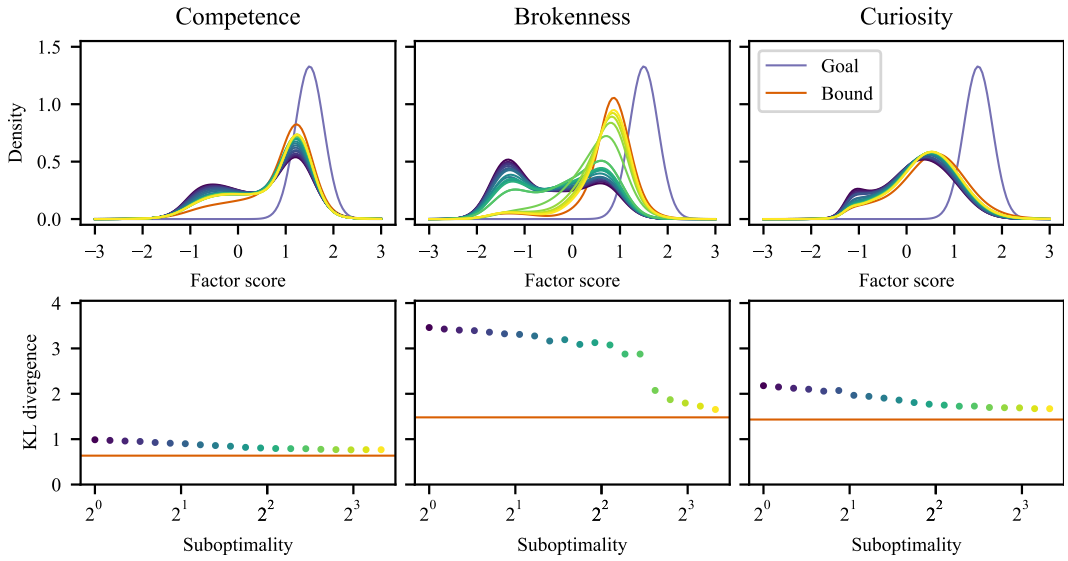
Template Insertion : A sequence of states is patched into the trajectory beginning at each s_i. We used a "straight" template, formed by taking the action used in s_{i-1} and repeating it twice, and a "U" template, formed by taking three actions in proceeding clockwise or counter-clockwise directions. The trajectory is reconnected with a shortest path to s_{i+1}. Both templates were directly motivated by participant feedback highlighting these patterns as contributing to the appearance of an organized principle to the robot's motion.

Collision Seeking : The trajectory is scanned and collision-avoiding actions are individually replaced with collision-causing actions if they are available from a state s_i. The trajectory is then reconnected with a shortest path to s_{i+1}. This sampler was motivated by participant responses highlighting collisions as contributing to their ratings of components of the brokenness score. These same modifications are generated by the "Action modification" sampler. We double sample these to reduce the chance that they are dropped in a subsequent subsampling step.

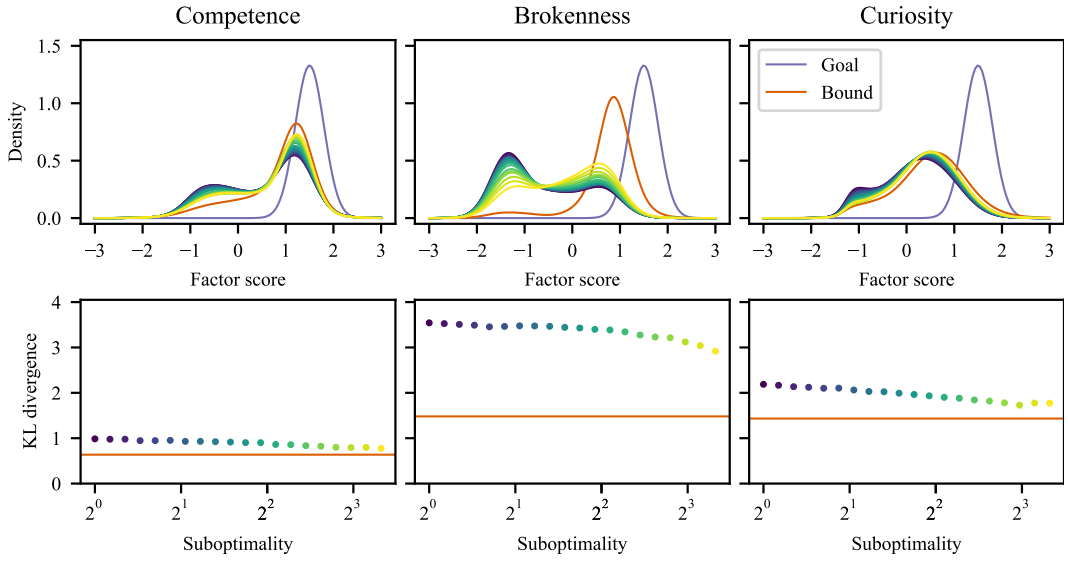
Overcoverage : A random section of the trajectory $s_i..s_j$ between 3 and 6 states long is selected uniformly at random and the trajectory is modified to backtrack this portion by inserting $reverse(s_i..s_j)$ followed by $s_i..s_j$ in place of s_j. This modification is motivated by participant feedback that highlighted redundant coverage as evidence of attentiveness, exploration or attention to detail.

These samplers, and a good initialization produced by solving the original task using a planner, make sampling-based greedy hill-descending search in the space of trajectories effective. As practical considerations, we discard sampled trajectories longer than 250 states to prevent the search from exploding, then we randomly subsample 250 of the modified trajectories, and we terminate the search once 750 steps have been taken.

As illustrated in Fig. 7, this search process is generally effective at making use of additional allowable suboptimality to produce trajectories that the model predicts to better elicit a desired attribution. Some non-monotonicity can be observed in one of the brokenness and one of the curiosity plots. As trajectories get longer, the pool of neighbors sampled is diluted with many more small changes and our subsampling step can lead the search to randomly discard more fruitful paths.



(a) Training environment (as used in Experiment I)



(b) Modified environment (as used in Experiment II)

Figure 7: The predicted distribution of factor scores for trajectories optimized under varying settings of w, represented as a ratio of the task-optimal cost. The optimization goals are configured as they are in our experiments, a Gaussian centered at 1.5 with standard deviation 0.3, and the plots show the marginal density for the factor under consideration. The companion plots show the KL divergence for each of the plotted distributions, providing a measure of closeness to the goal distribution for each generated trajectory. The bound line represents the distribution and KL divergence that are predicted for a feature vector obtained by optimizing the features directly, as opposed to optimizing in the space of trajectories. In general, these features may not be realizable with a trajectory that obeys the domain constraints, so the predicted distribution represents an upper bound.

E Statistical Results

We conducted one-sample Kolomogorov-Smirnov tests comparing the empirical distribution of factor scores elicited from participants against the distribution predicted by our model. To account for the increased likelihood of Type-I errors due to multiple testing, the Holm-Sidak adjustment was applied to the resulting p values. The results of the tests are provided in Table 5.

	Experiment I						Experiment II					
	Competence		Brokenness		Curiosity		Competence		Brokenness		Curiosity	
	$D(24)$	p	$D(24)$	p	$D(24)$	p	$D(24)$	p	$D(24)$	p	$D(24)$	p
1x	.381	**.014**	.292	.255	.233	.659	.343	.059	.247	.464	.319	.143
2x	.190	.789	.185	.789	.120	.879	.309	.145	.151	.836	.299	.206
4x	.275	.334	.232	.659	.144	.879	.182	.738	.256	.446	.254	.464
12x	.277	.334	.217	.695	.198	.789	.246	.464	.211	.601	.143	.836

Table 5: Results of Kolmogorov-Smirnov tests comparing predicted and observed distributions for each condition

Experiment I						Experiment II					
Competence		Brokenness		Curiosity		Competence		Brokenness		Curiosity	
τ_b	p	τ_b	p	τ_b	p	τ_b	p	τ_b	p	τ_b	p
-.261	**.001**	.402	**<.001**	-.069	.370	-.175	**.045**	.351	**<.001**	-.033	.675

Table 6: Results of Kendall's tau-b correlation between optimization parameter w and factor score

For Experiment I, the tests failed to indicate a significant difference between the distributions for all but the Competence-1x condition, suggesting that there is insufficient evidence to support inequivalence between predicted and observed distributions. For Experiment II, the tests failed to indicate a significant difference between the distributions for all conditions, suggesting that there is insufficient evidence to support inequivalence between predicted and observed distributions.

Kendall's tau-b correlations were run to determine the relationship between the allowable suboptimality of a generated trajectory (1, 2, 4 or 12) and the observed factor scores for the attribution under study. The correlation coefficients and p values are reported in Table 6. The results indicate strong positive correlations between the allowable suboptimality and brokenness factor scores, suggesting that the trajectory generation method was effective at eliciting progressively higher factor scores. However, tests for the competence conditions indicated a moderate negative correlation, and tests for curiosity conditions were not significant.