Human-Centered Design of Robot Explanations

Rosario Scalise

CMU-RI-TR-17-12

Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

As robots perform tasks in human-occupied environments and especially when those tasks are collaborative, people are increasingly interested in understanding the robots' behaviors. One approach to improving understandability is to enable robots to directly explain their behaviors, either proactively, or in response to a query. Given the many possible ways that explanations could be generated, our goal is to understand how people explain actions to other people.

We achieve this through a crowdsourcing approach that captures human explanations and then ranks those explanations using measures of clarity and generalizability. We then use those findings to propose human-centered design principles for robots' explanations to follow the human patterns that were ranked highest.

Our first set of studies focus on the natural language humans use when referring to blocks on a tabletop. We draw parallels to findings from psychology literature and present statistics about what composes a clear natural language reference. Our second set of studies focus on finding characteristics of robot trajectory navigation demonstrations that convey the most information about a robot's underlying objective function. We develop a theory of critical points along a trajectory and summarize how including these points in demonstrations affects human understanding of a robot's behaviors. Given our human-centered principles for explanation, we propose both perception and natural language algorithms to allow real robots to generate these explanations automatically.

Contents

1 Introduction

Robots are finding their way into many aspects of society. As this trend continues, interacting with robots will become a daily occurrence. In many instances, people will not only passively observe the robots, but also collaborate with them to jointly achieve a task. Just as we seek to make human interaction more fluid by striving for behaviors and strategies to aid in expressing our intentions, we should accommodate the same capabilities in our robot partners. Our task is to address the problem of robot understandability.

One approach in doing so is to enable a robot to explain its own behaviors. However, this can be challenging due to the vast space of parameters to consider when designing a system that can achieve this. In this work, we explore robot explanation methods with an emphasis on human-centered design principles as a way to focus our design. We observe human strategies for explanation through crowdsourcing and provide design guidelines based upon examples that other humans most easily understand.

Our work investigates two strategies. We first examine best practices in enabling a robot to specify its goals in natural language in a table-top manipulation task setting. We then examine what qualities make robot trajectory demonstrations effective at conveying a robot's underlying objective function. These two explanation methods are complementary to each other in that the former aims to specify the robot's hard constraints (e.g. its goals) where the latter aims to elucidate its soft constraints (e.g. the rules it must adhere to while trying to achieve its specified goal).

We contribute a framework for collecting and evaluating crowdsourced datasets of human examples. We first show that humans can more effectively interpret natural language which specifies an object within a scene when it has few spatial words in it and avoids language that is dependent on perspective. We also show that humans reference particular visual landmarks in scenes in order to establish groundings which tie language to the environment. Our observations coincide well with the human-psychology literature and we draw parallels wherever possible. Finally, we translate our findings into guidelines to be used when designing a robot explanation system which can specify its goal unambiguously.

We use the same 'collect and evaluate' framework to glean how humans perceive different types of robot trajectory demonstrations. We examine the trajectories that lead people to the most accurate understanding of the robot's underlying objective function and find that they share multiple characteristics. We define these characteristics as 'critical points' and systematically test how each type of critical point affects human understanding. We find that these points should be employed differently when the robot has a preference for traveling through certain states on its way to a goal state vs. when the robot does not have a preference and is purely seeking the goal. We succinctly summarize these findings and report them as a set of guidelines for how a robot can convey its objective function through its choices of actions in an effective manner.

2 Related Work

2.1 Intelligibility of AI Systems

The concept of intelligibility in AI systems was introduced by [5] and is defined as a systems capability in explaining its behavior. It has been cited as an important requirement for maintaining user satisfaction and usability in context-aware applications [20].

Within context-aware computing, there has been work in generating explanations to improve intelligibility [55] which answer questions like 'what did the system do?', 'why did it do X?', 'why not Y?', and 'what if Z?'. Improving intelligibility helps novice end-users understand a system and consequentially makes it easier for the user to trust the system [15, 55].

2.2 Explaining to Humans

[61] reactive dialog system where a robot explains itself to humans and get feedbacks.

[59] subsumptions in Description Logics

[34] explain solutions in problem-solvers and automated theorem provers.

[78] first-order logic based on a partial order of actions and causal links between actions in a hybrid planning system.

[41] explain the reason behind an optimal action in MDP

[84] diagnosis system which uses a sequence of actions to explain the system state

[29, 31] explain the causes of planning failures

[31] generate analysis of unsolvable tasks to help engineers model new planning application by trying operations.

[60] enable a robot to explain the causes of failures by proactively modeling the exogenous events, which could help planning agents better predict future states and replan.

[42] minimal revision problem for finding the "closest" satisfiable specification to the initial user specification when a robot fails to accomplish a task.

[10] explain the evolution of planning domain models by highlighting potential plan flaws, asking clarifying questions, and explaining model drift.

[50] debrief to human commanders by explaining its decision-making

[15] multi-model explanation generation as a model reconciliation problem,

[15] explanation = a prolonged interaction between humans and robots.

[95] plan explicability and predictability

[49] explicability

[12, 46, 47] investigate the effects of explanations of system reasonings considering end user personalization on the user mental models of the system.

[12] explanations on trust and reliance in clinical decision support systems.

[77] convert visualization to verbalization in natural language to describe mobile robot navigation experiences.

[68] further refine and predict robot verbalization space which is used to cover the variability of utterances that the robot may use to narrate its experience to different humans through continued dialog.

[88, 89] generate comments by detecting many events recognized from game controller and robot vision. [1, 56] describes team actions in surveillance and in training of military or sport teams through a spatio-temporal correlated pattern of movement modeled by a Hidden Markov Model.

[6] a route navigation dialogue system people can understand and follow.

[30] develops a turn-taking dialogue in human-robot social interaction.

robots understand and anticipate human intentions and plans [82] robot learn motion constraints from humans through natural language [35, 92]. Robot translate verbal commands from human commanders into multi-objective path planning problems [92]

Plan recognition problem [73]

One approach is to assume that all possible plans are given by a plan library, and recognize the plans through parsing algorithms with a grammar [27, 72], making inference in a Bayesian network [11, 54], and specialized procedures [3, 36, 40, 51].

Another approach is more generative with an agent action model and a set of possible intentions without plan libraries. [73, 74] use classical planning model to model agent actions assuming deterministic actions, [4] uses Markov Decision Process (MDP) to model agent actions assuming stochastic actions, and [75] uses Partially Observable MDP (POMDP) to model agent actions and infer a probability distribution over the possible agent intentions.

2.3 Plan Synthesizing

Robots optimize for human safety and comfort [82].

Safety minimize collision [82].

designing soft robots [7], developing distributed macro-mini actuation approach [96], tolerating system fault [57], minimizing danger index [48], minimizing danger index [64], minimizing danger index [37], control trajectory and velocity [45].

Comfort human "physical" safety vs "mental" safety, optimizing the psychological influence of robot motions on human partners [65]. optimize for psychological influence of robot motions on human partners [65].

optimizing wheelchair motion for human comfort [76], plan motions optimizing for the comfort distance between robots and humans [17, 26], enable socially interactive robots to imitate human motions in navigation [2, 66, 67, 82], optimize for motion constraints from the adverbs based on human input in natural langauge [91–93] and palette-based user interface [80, 81], and topological constraints specified by humans in natural language [94], plan tasks while avoiding conflicts with humans on shared resources [16, 18, 85], collaboratively execute shared plans with humans [79], develop proactive help and anticipatory actions to improve the fluency of human-robot collaboration [14, 32, 33].

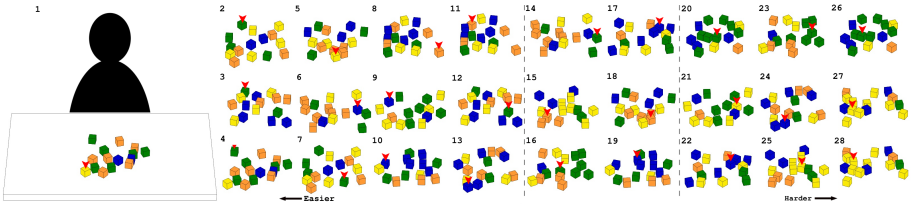


Figure 1: Scenes used to elicit spatial references. Online participants were asked to write how they would instruct the silhouetted figure to pick up the block indicated with the red arrow. For each participant, the silhouette was either referred to as a robot or a human partner. The block configurations on the left were rated as the easiest to describe, while the configurations on the right were the most difficult.

3 Design Guidelines: Natural Language for Robot Manipulation

3.1 Introduction

As people and robots collaborate more frequently on spatial tasks such as furniture assembly [43], warehouse automation [22], or meal serving [38], they need to communicate clearly about objects in their environment. In order to do this, people use a combination of visual features and spatial references. In the sentence "The red cup on the right", 'red' is a visual feature and 'right' is a spatial reference.

There is a long line of research in robotics related to communicating about spatial references like 'furthest to the right','near the back', and 'closest' for navigation task [9,35,44,58,83,86,87]. However, there are fewer studies involving the communication of spatial references for tabletop or assembly tasks [8]. A common theme in the space of tabletop manipulation tasks is clutter which we view as many potential objects to reason about. See Fig. 1

A cluttered table introduces the problem of *object uniqueness* where if there are two objects which are identified in the same manner (e.g. the red cup among two red cups), we are left with an ambiguity. One possible solution to this is to utilize *spatial references* which allow the use of spatial properties to establish a grounding or certainty about the semantic relationship between two entities.

However, even with the use of spatial references, it is still possible to encounter additional ambiguity which originates from the reference frame. Humans often use perspective to resolve this ambiguity as in the example 'the red cup on your right'. Often times, in tabletop scenarios, the person giving instructions will be situated across the table from their partner and thus will have a different perspective. Therefore, robots that collaborate with humans in tabletop tasks have to both understand and generate *spatial language* and *perspective* when interacting with their human partners. We investigate these key components by collecting a corpus of natural language instructions and analyzing them with our goal of clear communication in mind.

We first conducted a study in which we asked participants to write instructions to either a robot or human partner sitting across the table to pick up an indicated block from the table as shown in Fig. 1. This task raises a perspective problem: does the

participant use the partner's perspective or their own perspective, if any? Blocks were not always uniquely identifiable, and so the task required participants to describe spatial relationships between objects as well. We analyze the instructions from participants for 1) language differences between instructing a human versus a robot partner, 2) trends in language for visual and spatial references, and 3) the perspective(s) participants use when instructing their partners.

To investigate the effect of perspective, we conducted a second study in which we presented new participants with the instructions from the first study and asked them to select the indicated block. We utilized the correct selection of the indicated block as an objective measure of clarity. In order to establish which instructions contained ambiguities (lack of clarity), we first manually coded the instructions for whether the reference perspective was unknown or explicit (participant's, partner's, or neither) and whether there were multiple blocks that could be selected based on the instruction. An unknown perspective implies the instruction is dependent on perspective, but it is not explicitly stated.

Results from the first study show that participants explicitly take the partner's perspective more frequently when they believe they are instructing a person rather than a robot. Additionally, we find that people use color most frequently to refer to a block, while block density (e.g. the cluster of green blocks), block patterns (e.g. lines of red blocks), and even certain precise quantitative terms (e.g. 2nd block to the left) are also widely used. Finally, people spend more time writing the instructions and rate their tasks as more challenging when their instructions require the use of more spatial references.

From the second study, we find that 58% of our collected instructions contain perspective-dependent spatial references. Of this 58% more than half fail to explicitly specify the perspectives. This results in participants taking longer amounts of time to process the instructions and lower accuracies in discerning the intended block. The other 42% of instructions contained perspective-independent spatial references. These instructions demonstrated quicker completion times and higher correct block selection accuracies. We conclude that it is beneficial for instructions to avoid perspective-dependent spatial references when possible.

3.2 Study 1: Collecting Language Examples

To understand how people describe spatial relationships between similar objects on a tabletop, we collected a broad corpus of spatial references generated by 100 online participants. We analyzed this corpus for the types of words participants used and the word choice across differences in perceived difficulty of providing a spatial reference.

3.2.1 Study design

To collect spatial references that represents tasks that required perspective taking as well as object grounding, we created a set of stimulus images. Each image represents a configuration with 15 simplified block objects in different colors (orange, yellow, green, or blue) on a table. (Fig. 1). We first generated 14 images of configuration independently, each of which included different visual features and spatial relationships,

such as a single block of one color, pairs of blocks closely placed, blocks separated from a cluster, and blocks within or near clusters of a single color. Then we placed red-arrow indicators above two different target blocks independently in each image and ended up with 14 pairs of configuration (28 images of configuration in total).

This stimulus design is chosen to elicit instructions that rely more on the visual and spatial arrangement of the blocks than their individual appearance for the purposes of human-robot interaction. In order to capture clear instructions for a potential partner, this task asked participants to instruct a hypothetical partner to pick up the indicated block as though that partner could not see the indication arrow. The partner (indicated by the silhouetted figure in the images) was seated across the table from the participant viewing the scene. This setup required participants to be clear about the target blocks and the perspectives where they were describing the blocks.

Prior work indicates that people communicate with robots differently from with other people [13, 25, 62]. Therefore, we varied whether participants were told that their partner (the silhouette) was human or robot. [1] Participants were randomly assigned to either the human or the robot condition, and this assignment was the same for every stimulus they saw. The stimuli were otherwise identical across conditions.

We analyze the results with respect to these hypotheses:

H1 People use different words when talking to human and robot. Specifically, people are *more verbose*, *more polite*, and use *more partner-based perspective words* to human partners than robot partners.

H2 The frequency of words used in all instructions correlates with the features used in visual search (*color, stereoscopic depth, line arrangement, curvature, intersection, and terminator* [90]).

H3 Subjective ratings of sentence difficulty correlate with the number of spatial references required to indicate the target blocks.

3.2.2 Study Procedure

We deployed our study through Amazon's Mechanical Turk[2]. Each participant was randomly assigned a partner condition (human vs robot) and 14 trials. In each trial, participants were presented with an image, like the one on the left side of Fig. 1, which was randomly chosen from the two predefined configurations in each of the 14 pairs of configuration 3.2.1. The participants then typed their instructions and rated the difficulty of describing that block on a 5-point scale. For each trial, we also collected the completion time. After completing 14 trials, participants were asked 1) if they followed any particular strategies when giving instructions, 2) how challenging the task was overall, and 3) for any additional comments they had about the task. Finally, we collected demographics such as age, gender, computer usage, handedness, primary language (English or not), and experience with robots.

[1] We did not change the visual appearance of the silhouette

[2] www.mturk.com

Type	P1	P2	Example
Participant Perspective	+	-	"the block that is to **my** rightest."
			"**my** left most blue block"
Partner Perspective	-	+	"the block on **your** left"
			"second from the right from **your** view"
Neither Perspective	-	-	"closest to you"
			"the top one in a triangle formation"
Unknown Perspective	?	?	"to the **left** of the yellow block"
			"the block that is on far **right**"

Table 1: Possible perspectives. (P1=Participant P2=Partner).

3.2.3 Metrics

We analyze the collected corpus for language features. To analyze the differences on word choice between human-partner group and robot-partner group (H1), we computed

- *word count* - number of words for each instruction,
- *politeness* - presence of the word "please" in each instruction,
- *perspective* - whether the instruction explicitly refers to participant's perspective (egocentric), partner's perspective (addressee-centered), neither perspective[3], or unknown perspective (instruction implicitly refer to some perspectives) (see Table 1 for details).[4] [52,53]

Word count and politeness were automatically extracted from the text. Perspective was manually coded by four raters who coded the same 10% of the data and iterated until high inter-rater reliability, measured by averaging the result of pairwise Cohen's κ tests. The average κ value for perspective was 0.85, indicating high inter-rater reliability. Once this reliability established, the four raters each processed one quarter of the remainder of the instructions.

To compare the features used in our collected instructions with visual search (H2), we classify words into categories adapted from visual search literature [90]. The categories are listed in Table 2 and presented in the order of *word frequency*, the number of instructions that contain words from the category divided by the size of the corpus.

To verify the correlation between perceived difficulty and the number of required spatial references (H3), we compare the subjective *difficulty rating* (Likert scale 1 (easy) to 5 (difficult)) to the following objective measures:

- *word count* - as computed for H1
- *spatial reference count* - as computed for H2

[3] *Neither Perspective* sentences only use perspective-independent directional information. For example, "closer to you" should be classified as neither perspective instead of partner perspective, because it contains a perspective-independent reference to a landmark, "you," but not perspective-dependent relationships such as "on my left" and "on your right".

[4] Object-centered perspective is not considered because blocks are all the same except color

Word Category	Description
Action	An action to perform
Object	An object in configuration
Color	Color of object
Ordering/Quantity	Ordering/Quantization of objects
Density	Concentration of objects (or lack of)
Pattern/Shape	A readily apparent formation
Orientation	The direction an object faces
Environmental	Reference to an object in the environment
Spatial Reference	Positional reference relating two things
Perspective	Explicitly indicates perspective

Table 2: Word categories and their brief descriptions

- *ordering and quantity word count* - as computed for H2
- *completion time* - the duration from when a participant loads a new stimulus to when the participant hits the submit button for his/her instruction.

3.2.4 Results

In the study, we recruited 120 participants and over-sampled 1680 instructions so that we could account for errors in data collection process and invalid responses. We remove 10 sentences (0.006%) that either do not refer to any blocks or are otherwise nonsensical. For consistent and organized analysis, we randomly select 1400 sentences from the remaining 1670 to ensure that each of the 28 configurations has exactly 50 instructions divided as evenly as possible between partner conditions. We analyze the 1400 sentences selected in this manner.

3.2.5 Hypothesis H1

To evaluate the different words people used when speaking to a robot or human partner (H1), we analyze the overall *word count*, number of *politeness* words, and *perspective* used between the two partner conditions.

To analyze word count, we conduct an independent-samples t-test comparing number of words in the sentences for the two partner conditions. There is no significant difference in the mean sentence length by partner type (human: $M = 14.90, SD = 7.8$, robot: $M = 14.35, SD = 7.1$), $t(1398) = -1.389, p = 0.179$.

To analyze politeness, we conduct a Chi-squared test of independence between partner type (human or robot) and politeness word (present or absent). There is a significant relationship between the two variables, $\chi^2(1) = 6.685, N = 1400, p = 0.01$. Though use of politeness words is rare overall (only 4.6% of all the sentences contain "please"), politeness words are used significantly more often in human-partner condition (6.1%) than robot-partner condition (3.2%).

Visual Feature	Count	Frequency
Color	1301	0.929
Ordering/Quantity	498	0.356
Density	456	0.326
Pattern/Shape	60	0.043
Orientation	1	0.001

Table 3: Visual feature frequencies and feature-included sentence counts over all 1400 sentences ranked from most to least frequent

To analyze perspective, we conduct a Chi-squared test of independence between partner type (human or robot) and perspective used (participant's, partner's, neither, or unknown). There is a significant relationship between the two variables, $\chi^2(3) = 13.142, n = 1400, p = 0.004$. Post-hoc analysis with a Bonferroni correction identify that the partner perspective is used significantly more frequently in human-partner condition (28.1% of sentences) than in robot-partner condition (20.6% of sentences), $p = 0.001$. No other significant differences are found. This result is aligned with the idea that people adapt to the robot's assumed linguistic and perceptual abilities when talking to a robot. [62].

Thus, H1 is partially supported: there is no difference in sentence length between human and robot conditions, but people use "please" more often and take partner's perspective more frequently when they believe they are instructing another human than instructing a robot.

3.2.6 Hypothesis H2

To address our belief regarding the correlations between the visual features in our collected instructions and visual search (H2), we analyze how frequently sentences contain visual search features.

A summary of the results are in Table 3.

First, a reference to color is used in nearly every sentence, since color is such a salient feature in our stimuli as well as in visual search. Next, although orientation is also strongly influential according to visual search literature, orientation is almost never referenced in our data. This is likely due to the fact that in our study, blocks have 4-way symmetry and are not oriented in any particular direction [90].

Without many other visual indicators, participants frequently referred to "dense" regions of one particular color or to shapes or patterns they saw in the blocks such as a "line of red blocks". These references are observed in the literature with less consistency than color and orientation are [90].

Finally, although ordering/quantity does not fit the paradigm of visual search [90] as well as the previously mentioned features did, these words are closely related to the concepts of pattern/shape and density. "The third block in the line" and "The second block from the cluster" are examples respectively. We find high occurrence of ordering/quantity words especially in relation to other visual search terms.

In summary, we find that the observed frequency of many categories of words in our

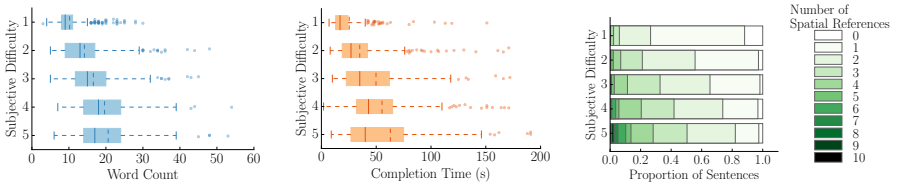


Figure 2: The effect of subjective difficulty on ratings and measures of the sentences, such as (a) word count, (b) completion time, and (c) number of spatial references.

corpus, including color, density, shape, and ordering/quantity, closely matched what we expected based upon the visual search literature [90].

3.2.7 Hypothesis H3

We evaluate the effect of perceived difficulty on word choice in each instruction (H3) by investigating the correlations between subjective rating of difficulty, overall word count, number of spatial references, number of order/quantity words, and completion time. We excluded any trials on which the participant did not provide a subjective rating of difficulty and two outlier trials for which the response times were greater than 10 minutes, which ended up with 1353 sentences.

Because we use ordinal measures in this evaluation (e.g. subjective difficulty is rated on a 5-point scale), we conduct a Spearman's rank-order correlation to determine the relationship among the five metrics identified. There are statistically significant correlations across all pairs of metrics ($p < 0.01$ for all, which accounts for multiple comparisons).

Table 4 details these correlations, and Fig. 2 visually displays some of them. Some of our key observations are:

1. As expected, there is a clear positive correlation (0.528) between word count and difficulty (Fig. 2a): easier scenes require fewer words to describe.

2. Also as expected, there is a clear positive correlation (0.508) between completion time and difficulty (Fig. 2b): harder scenes require more time.

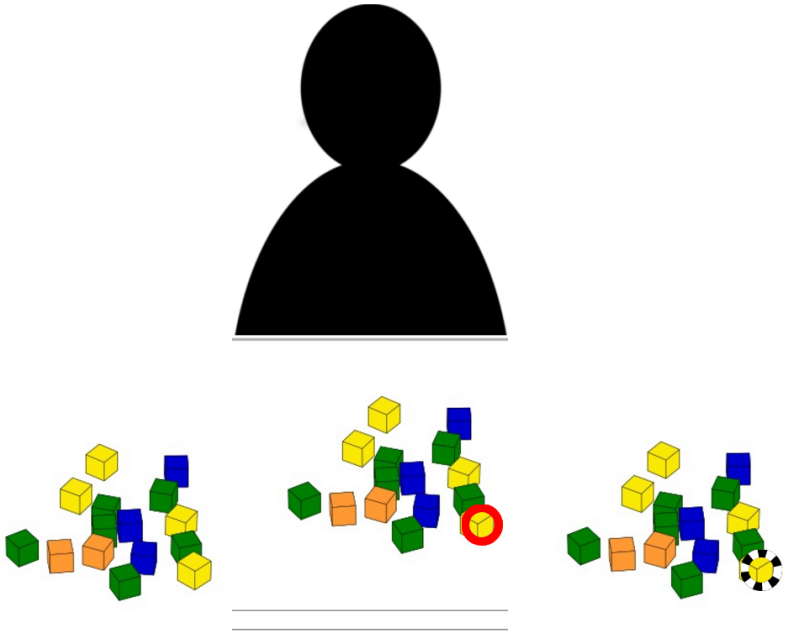
3. Interestingly, easier rated tasks generally require fewer spatial references (Fig. 2c): more spatial references in a sentence imply a greater depth of search to find the correct answer.

These findings confirm that subjective ratings of sentence difficulty are strongly correlated with the number of spatial references required to indicate the target block.

We conclude that participants are more polite and use partner's perspective more frequently when instructing a human partner than a robot partner. Additionally, the words used in the instructions are in line with the words used by participants when helping partners perform visual search. Finally, there are strong correlations between subjective rating of difficulty with all of our objective measures. However, we are

	Diff.	Word Count	Spatial Ref.	O/Q Word	Compl. Time
Difficulty	—	0.528	0.213	0.338	0.508
Word Count	0.528	—	0.416	0.425	0.682
Spatial Reference	0.213	0.416	—	0.082	0.262
Order/Quantity Word	0.338	0.425	0.082	—	0.350
Completion Time	0.508	0.682	0.262	0.350	—

Table 4: Spearman's rho correlations of sentence features and scene difficulty evaluations. All correlations are statistically significant with $p < 0.01$.



(a) (b) (c)

Figure 3: (a) In Study 2 (Sec. 3.3), we removed the red indication arrow. (b) As participants move their mouse over the image, a red circle will appear over the blocks to show them which block they could possibly select. (c) When they click on the block, a black and white checkered circle will appear around the selected block.

mostly interested in whether these collected instructions are clear enough for partners to understand. Our second study is aimed at analyzing the corpus from Study 1 for clarity.

3.3 Study 2: Evaluating Language for Clarity

To study the principles of clear spatial references in human robot collaboration, we need to validate the clarity of the instructions obtained in Study 1 (Sec. 3.2). First, we

manually coded the instructions in terms of two criteria (perspectives had already been coded in Study 1 (Sec. 3.2)):

- *block ambiguity* - the number of blocks that people could possibly identify from the image based on the given instruction.
- *perspective* - whether there is an explicitly stated perspective provided in the instructions.

Subsequently, we ran a follow up study to empirically measure the clarity of the sentences. In this second study, participants were presented with the stimuli from Study 1 (Sec. 3.2) (without red indication arrows) alongside the corresponding block descriptions from Study 1 (Sec. 3.2), and were asked to click on the indicated blocks. We collected responses from ten participants for each instruction from Study 1 (Sec. 3.2).

3.3.1 Coding Instructions for Clarity

We manually code each of the instructions from Study 1 (Sec. 3.2) for perspective and general block ambiguity. The coding measures, inter-rater reliability scores, and preliminary findings are described next.

3.3.2 Perspective

As described in Sec. 3.2.3 and Table 1, all sentences are labeled with perspective information. Among all the 1400 sentences, 454 (32.4%) sentences use unknown perspective, 339 (24.2%) sentences use partner perspective, 15 (1.07%) sentences use participant perspective, and 589 (42.1%) sentences use neither perspective.

3.3.3 Block Ambiguity

Block ambiguity is the number of blocks this sentence could possibly apply to. For our definition, no inferences are allowed when determining block ambiguity. Every detail which could possibly lead to ambiguity should be considered and expanded to different referred blocks. For example, the spatial relation "surrounded" could mean either partially or fully surrounded, which makes the sentence "the block that is surrounded by three blocks" potentially ambiguous. Unknown perspective could also lead to block ambiguity if different blocks are identified under the assumption of different perspectives.

We manually code each of the instructions from Study 1 (Sec. 3.2) for "high" or "low" block ambiguity. If a sentence could refer to only one single block in the scene, it is rated as "low" ambiguity. Otherwise, it is rated as "high" ambiguity. We use the same process as in Sec. 3.2.3 to establish inter=rater reliability. On 10% of the data, the average Cohen's κ for the four raters is 0.68, indicating high rater agreement. Each rater subsequently code one quarter of the remaining data.

Among all the 1400 sentences coded, 895 (63.9%) sentences are not block ambiguous with only one block being referred to, while 492 (36.1%) sentences possibly refer to more than one block.

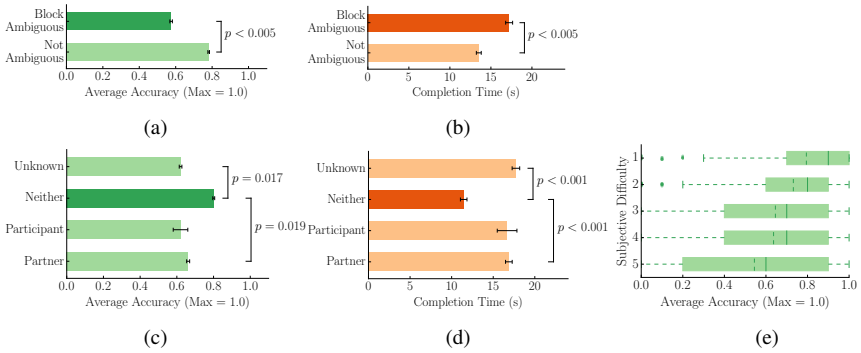


Figure 4: The effect of block ambiguity on (a) average selection accuracy and (b) completion time. The effect of perspective on average selection (c) accuracy and (d) completion time. The effect of the subjective participant ratings of difficulty from Study 1 (Sec. 3.2) on (e) average selection accuracy from Study 2 (Sec. 3.3).

3.3.4 Online Study Design and Procedure

As mentioned above, the goal of the second study is to investigate the clarity of spatial instructions, which will guide us through the future research on robot-generated instructions. In this online study, new Amazon Mechanical Turk participants were shown 40 configurations random chosen from the pool of 28 configurations generated in Study 1 (Sec. 3.2). Each configuration was presented alongside one of the corresponding instructions from Study 1 (Sec. 3.2) corpus. We would make sure that the clarity of all the collected instructions in Study 1 (Sec. 3.2) were evaluated here. Then the participants were asked to click on the block that best matched each instruction. As they moved their mouse over the image, a red circle appeared over the blocks to show them which block they would be selecting (Fig. 3b). When they clicked on the block, a black and white checkered circle would appear around the selected block (Fig. 3c). Continuing to move the mouse would present a red circle on those blocks which the participants could then click on to change their answer. Then we measured the participant's accuracy at selecting the indicated block.

We compute the following metrics for Study 2:

- *Final Answer* - whether a participant picks the correct block
- *Accuracy* - average over 10 participants of *final answer* for each instruction
- *Completion Time* - duration from moment when the page finishes loading to the moment when a participant clicks the next button to proceed.

Based on our ambiguity measures and the results from Study 2, we hypothesize that:

H4 *Block ambiguous* sentences will take participants in Study 2 **more time** and participants will be **less accurate** in discerning the referred block.

H5 Sentences with *unknown perspective* will take participants in Study 2 **more time** and they will be **less accurate** in discerning the referred block. Conversely, sentences with *neither perspective* will take **less time** and participants will be **more accurate** in discerning the referred block.

3.4 Results

We collect the responses from 356 participants and randomly select 10 responses for each of the 1400 sentences from Study 1 (Sec. 3.2). We evaluate the participant performance in Study 2 on the set of sentences from Study 1 (Sec. 3.2) by measuring their accuracy and completion time as described above. We also compare the objective accuracy measure to our manually-coded block ambiguity and perspective taking.

3.4.1 Hypothesis H4

First, we investigate block ambiguity by conducting an independent samples t-test measuring the effect of block ambiguity (low or high) on accuracy (Fig. 4a) and completion time (Fig. 4b). There are significant results for both accuracy ($t(1398) = 13.888, p < 0.005$) and completion time ($t(1398) = -5.983, p < 0.005$). Accuracy is lower and completion time is higher on sentences that contain ambiguous block references (H4). These results confirm that block ambiguous statements take longer amounts of time for participants to process and participants are less accurate in discerning the referred block.

3.4.2 Hypothesis H5

Next, we analyze perspective taking by conducting a one-way ANOVA measuring the effect of perspective type (participant, partner, neither, or unknown) on accuracy (Fig. 4c) and completion time (Fig. 4d). Perspective type has a significant effect for both accuracy ($F(3, 1396) = 43.655, p < 0.005$) and completion time ($F(3, 1396) = 34.607, p < 0.005$). Sentences that use neither perspective have higher accuracies ($M = 0.802, SD = 0.240$) than sentences that use partner ($M = 0.662, SD = .278, p = 0.019$) or unknown ($M = 0.619, SD = 0.307, p = 0.017$) perspective (H5). Similarly, average completion time is lower for sentences that use neither perspective ($M = 11.418s, SD = 10.56$) than partner ($M = 16.881, SD = 9.81, p < 0.001$) or unknown ($M = 17.756, SD = 12.03, p < 0.001$) perspective (H5). No other significant differences are found. These results confirm that neither perspective statements take shorter amounts of time for participants to process and participants are more accurate in discerning the referred block. At the same time, unknown perspective statements take participants longer time and participants are less accurate.

Additionally, we observe that participants in Study 2 have lower accuracy on sentences that participants in Study 1 (Sec. 3.2) label as more difficult (Fig. 4e). This result is not surprising as participants who have trouble writing a clear sentence would likely rate the task as difficult.

We conclude that hypotheses 4 and 5 are both supported. Block ambiguity and unknown perspective are both correlated with higher completion times and lower ac-

curacies. The type of perspective in the sentence has a significant effect on accuracy: when the instructions are written in neither perspective, participants in Study 2 have higher accuracy than any of the other perspectives.

3.5 Discussion

Keeping the goal of seemless human-robot collaboration in a tabletop manipulation setting in mind, we find the results from this first step forward quite encouraging. We created a corpus of natural language when specifying objects in a potentially ambiguous setting. We identified a cognitive process which plays a significant role in the formation of these specifying descriptions. We defined metrics to aid in scoring the optimality of a description. We designed an evaluation process based on these metrics. And finally, we performed an initial, yet broad, analysis on our corpus that was able to uncover a handful of insights. We will discuss a few of these insights in the following section.

In analyzing the corpus, we discovered that participants generally followed one of three approaches when writing instructions: (1) a *natural* approach where they used embedded clauses linked together by words indicating spatial relationships such as in the instruction "Pick up the yellow block directly to the left of the rightmost blue block.", (2) an *algorithmic* approach, which a majority of the users employed, where they partitioned their description in stages reflecting a visual search process such as in the instruction "There is an orange block on the right side of the table. Next to this orange block is a yellow block. Please pick up the yellow block touching the yellow block", (3) an *active language* approach where they provided instructions asking the partner to move their arms (usually) in a certain way so as to grasp the desired object such as in the instruction "Stand up, reach out over the table, and grab the yellow block that is touching the blue block closest to me.". In certain instructions, the participant would even offer active guidance (which is of course not an option in a one shot response written in a web form).

Among the three, the algorithmic approach is often the clearest but feels less natural. We believe that these observations about instruction approach types will lend themselves well to further investigation on user instruction preferences. For example, some users might prefer to give algorithmic descriptions which iteratively reduce ambiguity as needed, while other users might prefer to utilize active language where they guide the robots motions via iterative movement-driven instruction.

Our findings suggest that sentence clarity suffers when there is either an ambiguity related to the number of blocks a sentence can specify or an ambiguity related to perspective. An interesting observation is the relationship between block ambiguity and perspective ambiguity. Because the process we used in coding the data did not exclude one from the others, it was highly possible that these two features were dependent although the Pearson correlation indicated the opposite ($r = -0.0287$). Perspective ambiguity will often results in block ambiguity, except in the case that there features in the instructions that are dominate enough to eliminate all the possible blocks aside from one. For example, in "It is the block all the way on the right side by itself", the perspective is unknown but only one block in the scene is identifiable since it is the only "by itself" candidate. In this case, we can reduce the instruction to "It is the block

by itself". On the other hand, block ambiguity does not always imply unknown perspective. For example, in "pick up the closest green block", although the perspective is neither, not unknown, there are actually multiple possible blocks inferred from the instruction due to ambiguity in the landmark being referred to (e.g. closest to what?).

Further, descriptions requiring perspective always seem to include terms like 'right', 'left', 'above' and 'below'. We shall classify these as directional relative spatial references. If establishing perspective proves to be difficult in a scenario, and a sentence can avoid using directional relative spatial references, the robot should prefer to avoid these kinds of descriptions. That is, if the robot is able to generate a description using our definition of 'neither' perspective, it should prefer to do so over other descriptive strategies.

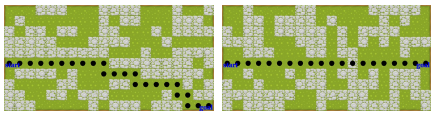
The intention of this work is to establish a baseline understanding of human preferences and behaviors when giving manipulation scenario instructions to a robot. We identify this one-shot language data analysis as a necessary step in laying the foundation for a truly interactive system which might take multiple rounds of input or ask questions to reduce uncertainty. Even without the element of active conversing, however, the results and insights we were able to extract are rather encouraging and have allowed us to establish effective grounding. We intend to gradually introduce interactivity in future works with varying approaches and modalities, and we believe the work we present in this paper provides an excellent initial benchmark.

4 Design Guidelines: Robot Trajectory Demonstrations

4.1 Introduction

As robots perform tasks in human-occupied environments, people who observe them form beliefs about their behaviors [28]. Without insight into the robot's objective function or other information about how the robot behaves, people must derive their expectations from only the robot's motion within the context of the environment. These beliefs guide peoples' understandings and expectations of the robots as well as their interactions. If a person cannot understand why a robot planned its trajectory – even a successful one – they may not be able to predict its trajectory in a new environment.

Prior work has focused on using robot motion to effectively convey robot capabilities and goals [21, 63]. In contrast, we focus on using robot motion to convey its own objective function and show that it prefers to navigate through states with particular features. Consider the trajectory shown in Fig. 5a. It appears that the robot does its best to avoid rocks while navigating to the goal, implying it has a preference for traversing grassy states over rocky states. However, this trajectory could have also been generated by a robot with an objective function that has no preference for either terrain type if it arbitrarily chose where to turn. Similarly, a person observing the robot in Fig. 5b may be unclear about whether the robot has no terrain preference or a strong preference for grass. A person who does not understand the robot's objective function could be confused in a new environment when it does not plan a new trajectory that matches their expectations. We are interested in producing robot motion trajectories that help people



(a) (b)

Figure 5: Many possible objective functions could generate these trajectories.

understand the robot's feature preferences and that improve their ability to generalize that behavior to new environments.

Based on the observation that people assign rational meaning to agent actions [19, 28, 39] we define two types of critical points in a trajectory – *inflection points* and *compromise points* – as points that are information-rich and convey information about the relationship between the planned trajectory and the features in the environment. Fig. 5a is an extreme example of how inflection points (i.e., changes in direction) may lead an observer to infer preference for grass because the trajectory traverses only that terrain feature. The single rock compromise point in Fig. 5b may similarly lead an observer to believe there is no preference for grass over rocks when in fact all alternative paths have more rocks and therefore a lower overall value.

Our goal is to determine, in detail, the roles these kinds of points play in trajectories that lead to good understanding of robot behavior. Towards this, we conducted a large-scale study to systematically examine how varying the critical points in trajectories affects peoples' understandings of robot behaviors. We generated trajectories through synthetic environments according to different robot behaviors and showed them to people via Amazon Mechanical Turk (AMT). We conducted a within-subjects study in which we varied the parameterizations of the robot's reward function as well as the combinations of critical points along each trajectory and asked people to specify their understandings as well as generalize new plans in different environments. We show that people understand and can generalize the robot's terrain preferences more accurately as the number of inflection points increases and compromise points decreases within trajectories. However, when a robot has no preference for terrain types, the addition of either type of critical point within a trajectory reduces a participant's understanding.

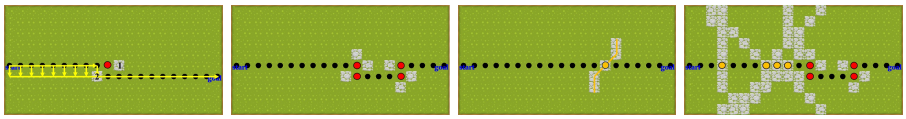
We conclude that our critical points in trajectories do provide observers more information about a robot's state preferences. A robot that can take these points into consideration while planning its trajectories can reduce observer uncertainty about its behavior while still acting optimally.

4.2 Problem Formulation

We formulate our robots' behaviors as a standard Markov Decision Process which is a tuple of the form: $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, R\}$.

This includes a set of world states $s \in \mathcal{S}$ with a single absorbing goal state $s_g \in \mathcal{S}$ and a set of robot actions $a \in \mathcal{A}$. The MDP has a deterministic state transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ and an immediate reward function $R : \mathcal{S} \rightarrow \mathbb{R}_+$. A robot behaves according to a deterministic policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$. The optimal policy is denoted as π^* and describes the policy that maximizes the overall reward.

A trajectory $\xi(s_0|\pi) \in \Xi$ is defined as a sequence of states $[s_0, s_1, s_2, ..., s_g]$ where



(a) (b) (c) (d)

Figure 6: (a) generating 1 inflection point (red dot) by placing rock tiles at state 1 and 2 (b) 4 inflection points (red dots) (c) generating 1 compromise point (orange dot) by building a frontier (orange line segments) (d) 4 compromise points (orange dots)

$\forall s_t \in \xi(s_0|\pi), \mathcal{T}(s_{t-1}, \pi(s_{t-1})) = s_t$. The total reward of ξ is $R_\xi(\xi) = \sum_{s_t \in \xi} R(s_t)$. An optimal trajectory ξ^* is yielded by following π^*.

To ensure there are no cycles in a trajectory, there is one and only one $s \in \mathcal{S}$ such that $R(s) \geq 0$.

4.2.1 Experimental Setup

As an example domain, we consider a gridworld representation of a park which has a single terrain feature such as grass or rock assigned to each state (tile) on the grid.

- State $s \in \mathcal{S}$ is defined as $s = (x, y)$
- Action a is a 4-connected movement where $a \in \mathcal{A} = \{\rightarrow, \uparrow, \downarrow, \leftarrow\}$
- We define $\phi : \mathcal{S} \rightarrow \mathbb{N}_+^3$ as a mapping from states to features. $\boldsymbol{\phi}(s) = [\mathbb{1}_{\text{goal}}(s), \mathbb{1}_{\text{grass}}(s), \mathbb{1}_{\text{rock}}(s)] \in \{0, 1\}^3$ subject to $\|\phi(s)\| = 1$, where each $\mathbb{1}(s)$ is an indicator function (e.g., $\mathbb{1}_{\text{grass}}(s) = 1$ if the tile type at s is grass and $\mathbb{1}_{\text{grass}}(s) = 0$ otherwise)
- We define $\mathcal{T}$ as a transition mapping with deterministic 4-connected movements within the gridworld.
- $\boldsymbol{\theta} \in \mathbb{R}^3$ are the weights for the feature vector $\boldsymbol{\phi}$. The reward for a state s with weights $\boldsymbol{\theta}$ is given by $R(\boldsymbol{\theta}, s) = \boldsymbol{\theta}^\mathsf{T} \boldsymbol{\phi}(s) \in \mathbb{R}$.
- When deriving the optimial policy, we break action ties with the ordering $[\rightarrow, \uparrow, \downarrow, \leftarrow]$.

4.3 Critical Points of Trajectories

Depending on a robot's functional objective, the trajectory it follows can vary significantly. We characterize the information-rich states and actions within a trajectory as *critical points*. Based on the rationality principle, we focus on two types of critical points – *inflection points* in which people assign meaning to changing direction and *compromise points* in which a robot traverses over states with different features. Although this set of characteristics is not exhaustive, we believe it provides an effective starting point in analyzing trajectories. We will demonstrate that critical points can be beneficial in guiding the observer's understanding of robot behavior, or they can be detrimental to an observer's understanding, confounding their beliefs and leading to misinterpretation.

4.3.1 Inflection Points

Inflection points are defined as $s_t \in \xi(s_0|\pi)$ where the robot changes its direction. In other words, inflection points are all points at which the robot's action is not identical to its prior action (i.e., $\pi(s_{t-1}) \neq \pi(s_t)$). In Fig. 6a, an inflection point is indicated by the red dot where the robot moves up. This change of behavior gives people information about the robot's aversion towards the rock tile annotated as 1. In our park environment, inflection points come in pairs (e.g., the two inflection points in Fig. 6a) because the robot typically resumes moving rightward after changing direction.

4.3.2 Compromise Points

Compromise points are defined as states $s_t \in \xi^*(s_0|\pi^*)$ in which the myopic reward of entering s_t is not the maximum obtainable from s_{t-1}, yet the total reward for the trajectory is maximized. In particular, $\exists\, a_{t-1} \in \mathcal{A}, a_{t-1} \neq \pi^*(s_{t-1}), \mathcal{T}(s_{t-1}, a_{t-1}) = s'_t.\ s.t.\ R(s'_t) > R(s_t)$, but $R_\xi(\xi^*(s'_t|\pi^*)) < R_\xi(\xi^*(s_t|\pi^*))$.

The trajectory in Fig. 6c contains one compromise point (orange dot). To reach the goal, the robot must traverse a terrain feature ($e.g.$ rock) which incurs a higher cost than another possible terrain feature (grass) accessible from the previous state. Any attempt to move around the rock frontier would result in lower total trajectory reward compared to the straight path over the one compromise point.

4.4 Generating Demonstrations

We develop a method for synthesizing trajectories through environments that demonstrate the robot's reward function $R(\boldsymbol{\theta}, s)$ by changing $\boldsymbol{\phi}$ by iteratively inserting inflection and compromise points into the trajectory ξ^*.

4.4.1 Inflection Points

To create an inflection point at $s_i \in \xi^*(s_0|\pi^*)$, we can decrease the reward of s_{i+1} which alters $\pi^*(s_i)$ to avoid s_{i+1}. In Fig. 6a, grass is preferred and has lower cost than rock. To create an inflection point at s_i indicated as the red dot, we place a rock terrain tile at s_{i+1} annotated as state $\mathbf{1}$.

One side effect of changing state $\mathbf{1}$ is that it might introduce multiple optimal policies yielding multiple optimal trajectories. The ambiguity of multiple optimal trajectories (or policies) can mislead people as it requires more complex reasoning to identify. One solution is to change some states to make all but one of the optimal trajectories suboptimal. We treat this as a set cover problem. Universe U is the set of all the available optimal trajectories $U = \{\xi | \xi = \xi^* \leftarrow \pi^*\}$. $\forall s \in \xi \in U$, we define $subset(s) \subset U$ to include all the optimal trajectories that go through s (i.e., $subset(s) = \{\xi | s \in \xi \in U\}$). The family set contains all the $subset(s)$ (i.e., $set = \{subset(s) | s \in \xi \in U\}\ s.t.\ \bigcup_{ss \in set} ss = U$).

Our goal is to find the minimum number of states that all but one of the optimal paths include (i.e., to find the minimal set $cover$ subject to $\bigcup_{cc \in cover} cc = U \backslash \xi^{**}$ where $\xi^{**} \in U$ is the only path $s.t.\ \forall cc \in cover, \xi^{**} \notin cc$). $\forall subset(s) \in cover$,

we can reduce $R(s)$ to make all the $\xi \in subset(s)$ sub-optimal and leave ξ^{**} the only optimal trajectory.

In Fig. 6a, there are 9 extra optimal trajectories available after changing states **1** (yellow arrows). By placing a rock terrain at state **2**, we could prevent the robot from moving downwards before reaching the red dot and make the trajectory indicated by black dots the only optimal trajectory. We generate 4 inflection points accordingly as shown in Fig. 6b.

4.4.2 Compromise Points

Similar to generating an inflection point, to generate a compromise point at $s_i \in \xi^*(s_0|\pi^*)$, we could decrease the reward of s_{i+1} $s.t.$ $R(s_{i+1}) < R_{\text{max}}$. But the difference is that now we want robots to keep $\pi^*(s_i)$ and head to s_{i+1} inevitably. Hence, we could decrease the rewards of a set of states in a neighboring area close to s_{i+1} to make it too costly for robots to detour around s_{i+1}. We could initiate the area as one state and iteratively increase its size until the new optimal trajectory passes through s_{i+1}. In each iteration, we could grow the area by making all the optimal trajectories which do not go through s_{i+1} become sub-optimal using the same technique we introduced in Sec. 4.4.1.

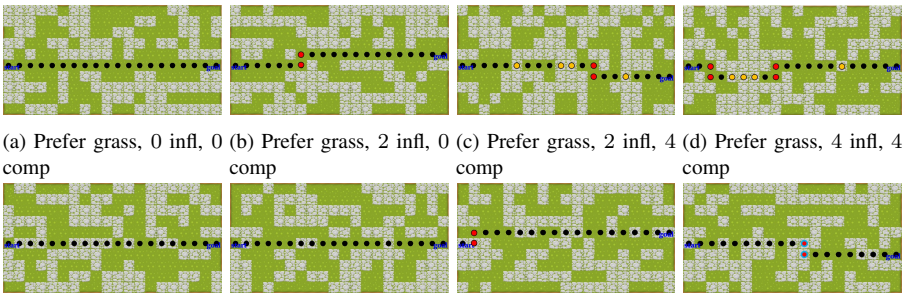
In Fig. 6c, to create an compromise point at s_i (orange dot), we can build a frontier of states filled with rock terrain from top to bottom across the entire map (orange frontier). This frontier with low reward will force the robot to pass through s_{i+1} (the black dot on the right next to the orange dot). In our implementation, we use cubic Bezier curves [23] randomly generated through De Casteljau's Algorithm [24] to represent natural-looking frontiers. We generate 4 compromise points accordingly as shown in Fig. 6d.

4.4.3 Extra Points

We uniformly distribute different ϕ's across our demonstration maps to so that all the maps are consistent with each other regarding the frequency of each feature. In our implementation, we ensure that each map contains 50% rocks and 50% grass adding complementary rock tiles to grass-dominant maps and vice versa. To make maps look natural, we place terrain types based on 2D Perlin noise [69–71]. Final maps and trajectories are shown in Fig. 7.

4.5 Empirical Evaluation

We ran a study to test the effects of trajectories with different critical points on human understanding of robot terrain preferences. We presented participants with 16 different maps of "parks" with rock and grass terrain features, containing trajectories starting from the left and traversing to the right side of the park. We manipulated the number of critical points within trajectories as well as the actual terrain preference demonstrated in each map, and measured each participant's ability to predict the robot's preferences in a within-subject study design.



(a) Prefer grass, 0 infl, 0 comp (b) Prefer grass, 2 infl, 0 comp (c) Prefer grass, 2 infl, 4 comp (d) Prefer grass, 4 infl, 4 comp

(e) No pref., 0 infl, 0 comp (f) No pref., 0 infl, 4 comp (g) No pref., 2 infl, different, 0 comp (h) No pref., 2 infl, same, 0 comp

Figure 7: Robot preference type, number of inflection points (red dots), inflection point configuration ("different" = red dots with black circles, "same" = red dots with blue circles), number of compromise points (orange dots) for demonstration examples

4.5.1 Independent Variables

We tested six terrain preference conditions and 10 no-preference conditions. The six preference conditions comprise all combinations of {0,2,4} inflection and {0,4} compromise points. The no-preference conditions are combinations of {0,2,4} inflection points, {0,4} compromise points, and {same, different} inflection point configurations[5].

Terrain Preferences. We compared trajectories through maps when there was a terrain feature preference versus when there was no preference between terrain features. We randomly selected half of the terrain preference conditions to prefer rock and half to prefer grass.

Inflection Points. Each demonstration trajectory had 0, 2, or 4 inflection points. Locations of the inflection points were randomly chosen along the path.

Compromise Points. We set the number of compromise points in each demonstration trajectory to be one of two values. When the reward function had preferences, these two values were {0, 4}. We were interested in observing the differences between having no compromise points versus having several points where the robot must ''make a compromise" (which we chose to be 20% of the total trajectory length). When the reward function had no preferences, compromises could not technically occur. Therefore, we arbitrarily assigned a ''simulated" preference and then divided the number of terrain features along the trajectory in the two levels: {50-50, 20-80}. The former level resulted in a trajectory where there was no preference illustrated by compromise points. The latter resulted in a trajectory where the robot simulated a compromise on 20% of the states.

Inflection Point Configuration. At each inflection point, there is a 'decision' corresponding to the change in direction. The robot's direction switches from continuing onto one tile (Fig. 8, B) to moving onto another tile (Fig. 8, C). We test whether human understanding changes if the terrain types of those tiles are the same (i.e., the robot

[5]When there are no inflection points, there are no inflection point configurations, hence there are 10 'no preference' maps instead of 12.

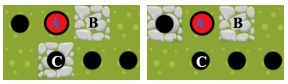


Figure 8: (left) an inflection point with 'same' configuration (right) an inflection point with 'different' configuration

chooses to turn from one grass tile to another grass tile) or different (i.e., the robot turns from a grass tile onto a rock tile). This condition is only tested when there is no preference in the terrain type.

4.5.2 Response Types

Sliders. We included a slider for each terrain feature type and labeled them {"Strongly Avoid", "Slightly Avoid", "Neutral", "Slightly Prefer", "Strongly Prefer"}. We asked participants to indicate the preference the robot had demonstrated for each terrain type using the sliders. Participants were free to place the sliders anywhere along the scale. We mapped their slider placements to a value between $[0, 1000]$, where 0 corresponds to "Strongly Avoid", 500 corresponds to "Neutral", and 1000 corresponds to "Strongly Prefer".

Text Free-Response. Participants were asked to explain the reasoning they believe the robot used as it planned its path through the map. Unlike the sliders, free response allows an unconstrained representation of the users' mental models of the robot behaviors. Due to space constraints, we do not present the results from the free response.

Drawing Trajectories. Last, we presented the participants with a new map (without a demonstration trajectory pre-drawn on it) and asked them to draw the trajectory they believed the robot would take if it were using the same reasoning to plan its new trajectory. Participants were required to start at a predefined point and could add 4-connected waypoints until reaching the goal position. Each map was generated to ensure it had a single optimal trajectory with respect to a fixed terrain preference. The maps were filled 50/50 with rock and grass tiles. In order to reduce the bias in our test maps, each participant received a randomized test map for each experimental condition. This measure allowed us to test participants' understanding of the robot's behaviors by comparing their drawn path to the optimal one.

Subjective Confidence. We asked participants to indicate on a 5-point Likert scale how confident they were that the trajectory they drew would be the one the robot would take.

4.5.3 Study Deployment

We recruited 90 participants via Amazon Mechanical Turk. We used a within-subjects design where each subject was shown the total 16 conditions (6+10) in the same order. This order was pre-determined to ensure that no three consecutive conditions had the same terrain preference, which allowed us to avoid users inferring incorrectly based on coincidental patterns. Upon completion of the study, we collected demographic information from participants, including their age, gender, occupation, primary language, and experience with robots, video games, and RC-cars. We also asked for general

comments as well as how difficult they found the tasks. Due to space constraints, these results are omitted.

4.6 Results

4.6.1 Dependent Variables

Our measures of accuracy in understanding robot preferences are based on the drawn trajectories, sliders, and subjective ratings of confidence.

The *optimality ratio* $= |\frac{\text{total cost of optimal trajectory}}{\text{total cost of drawn trajectory}}| \in [0, 1]$. As people understand the robot reward function more accurately, the optimality ratio increases.

We assume that people use the distance between the grass and rock slider placements to indicate their certainty about inferring the robot preferences. We map the distance between the grass and rock slider placements to *preference range* $\in [0, 2000]$. A value of 0 corresponds to the user inferring no preference between the grass and rock terrains while a value of 2000 corresponds to the user inferring a difference with a high certainty, regardless of what the robot actually prefers.

We use *subjective confidence* $\in \{1, 2, 3, 4, 5\}$ to represent the user's self-reported confidence in understanding robot reasoning, with higher values indicating more confidence.

4.6.2 Hypotheses

H1 Preference demonstrations: increasing the number of inflection points will increase optimality ratio, preference range, and subjective confidence.

H2 Preference demonstrations: increasing the number of compromise points will decrease optimality ratio, preference range, and subjective confidence.

H3 No preference demonstration: increasing the number of inflection points will decrease optimality ratio, preference range, and subjective confidence.

H4 No preference demonstration: increasing the number of compromise points will decrease optimality ratio, preference range, and subjective confidence.

H5 No preference demonstration: the optimality ratio, preference range, and subjective confidence are lower when each inflection point has a different configuration than when each inflection point has the same configuration.

4.7 Results

4.7.1 Preference

Optimality Ratio. We use a two-way repeated measures ANOVA to find the effect of inflection points and compromise points on optimality ratio (Table **??**).

The number of inflection points has a significant effect on the optimality ratio ($F(2, 178) = 46.159, p < 0.001$). Post hoc analysis with a Bonferroni adjustment reveals that the optimality ratio is significantly increased from 0 to 2 ($p < 0.001$) and

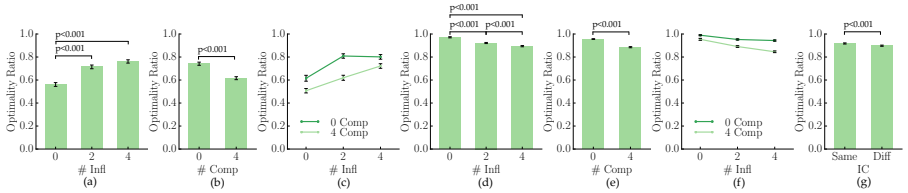


Figure 9: When there is a preference, optimality ratio vs (a) the number of inflection points (b) the number of compromise points (c) the interaction between the number of inflection points and compromise points. When there is no preference, optimality ratio vs (d) the number of inflection points (e) the number of compromise points (f) the interaction between the number of inflection points and compromise points (g) inflection point configuration

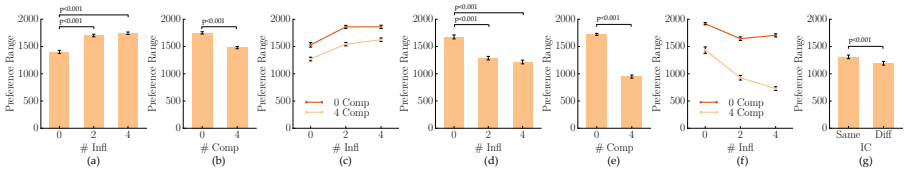


Figure 10: When there is a preference, preference range vs (a) the number of inflection points (b) the number of compromise points (c) the interaction between the number of inflection points and compromise points. When there is no preference, preference range vs (d) the number of inflection points (e) the number of compromise points (f) the interaction between the number of inflection points and compromise points (g) inflection point configuration

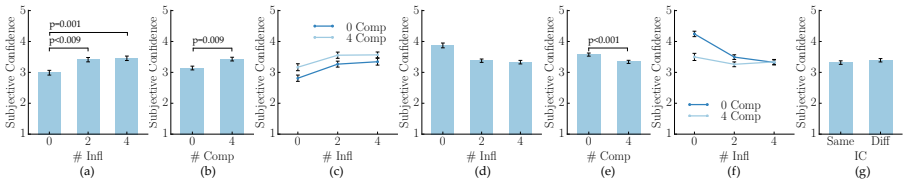


Figure 11: When there is a preference, subjective confidence vs (a) the number of inflection points (b) the number of compromise points (c) the interaction between the number of inflection points and compromise points. When there is no preference, subjective confidence vs (d) the number of inflection points (e) the number of compromise points (f) the interaction between the number of inflection points and compromise points (g) inflection point configuration

from 0 to 4 ($p < 0.001$), but not from 2 to 4 inflection points ($p = 0.052$), though it is close (Fig. 9a). This suggests that inflection points help users understand robot preferences. For example, it is easier for people to understand that the robot prefers grass over rock terrains by looking at Fig. 7b than Fig. 7a. Additionally, in these maps there is little benefit to showing more than 2 inflection points. For example, it is more difficult for people to understand that the robot prefers grass over rock terrains by looking at Fig. 7b than Fig. 7d although Fig. 7d has more inflection points. The first part of H1 is supported.

Increasing compromise points from 0 to 4 significantly decreases the optimality ratio ($F(1, 89) = 74.476, p < 0.001$) (Fig. 9b). This suggests that path entropy hinders people's understanding of robot preferences. For example, it is easier for people to understand that the robot prefers grass over rock terrains by looking at Fig. 7b than Fig. 7c. The first part of H2 is supported.

There is a significant interaction between the numbers of inflection and compro-

mise points on optimality ratio ($F(2, 178) = 5.291, p = 0.006$). When there are no compromise points, there is no significant difference between 2 and 4 inflection points ($p = 0.730$). However, when there are 4 compromise points, optimality ratio is significantly increased from 2 to 4 inflection points ($p = 0.001$) (Fig. 9c). This indicates that as the number of compromise points increases, people need more inflection points to mitigate their confusion about the compromise points. For example, it is easier for people to understand that the robot prefers grass over rock terrains by looking at Fig. 7d than Fig. 7c.

Preference Range. We used a two-way repeated measures ANOVA to determine the effects of inflection points and compromise points on preference range (Table **??**). The number of inflection points has a significant effect on preference range ($F(2, 178) = 65.759, p < 0.001$). A post hoc analysis with a Bonferroni adjustment reveals that the preference range is significantly increased from 0 to 2 ($p < 0.001$) and from 0 to 4 ($p < 0.001$), but not from 2 to 4 ($p = 0.385$) inflection points (Fig. 10a). This suggests that more inflection points lead to greater certainty about the robot's preference. Similar to optimality ratio, increasing beyond 2 inflection points does not improve preference range. The second part of H1 is supported. Preference range is also significantly decreased from 0 to 4 compromise points ($F(1, 89) = 91.050, p < 0.001$) (Fig. 10b). The second part of H2 is supported. There are no other significant effects on preference range.

Subjective Confidence. To measure the effect of inflection and compromise points on the Likert scale responses for subjective confidence, we ran a generalized ordinal logistic model and estimated the model parameters through a generalized estimating equation (GEE) with AR(1) covariance structure (Table **??**). Subjective confidence significantly increased from 0 to 2 ($p = 0.009$) and from 0 to 4 ($p = 0.001$), but not from 2 to 4 ($p = 0.907$) inflection points (Fig. 11a). This suggests that inflection points help people feel more confident about their evaluations, but that increasing beyond 2 inflection points does not necessarily lead to more confidence. The third part of H1 is supported. Subjective confidence is significantly increased from 0 to 4 compromise points ($p = 0.009$) (Fig. 11b). This suggests that path entropy decreases users' feelings of confidence in their evaluations. Interestingly, the third part of H2 is rejected. There are no other significant effects for subjective confidence.

4.7.2 No Preference

Analysis for no preference maps follows the analysis for preference maps above. Results for inflection point configuration are only available for demonstrations with 2 or 4 inflection points, since 0 inflection points mean there cannot be inflection point configurations.

Optimality Ratio. We conducted a three-way repeated measures ANOVA to determine the effect of inflection points, compromise points, and inflection point configuration on optimality ratio (Table **??**).

Number of inflection points significant affects optimality ratio ($F(2, 178) = 42.050, p < 0.001$). Post hoc analysis with a Bonferroni adjustment reveals that optimality ratio is significantly decreased from 0 to 2 ($p < 0.001$), from 0 to 4 ($p < 0.001$), and from 2 to 4 ($p < 0.001$) inflection points (Fig. 9d). This suggests that people's ability to iden-

tify the robot's true preferences continues to decrease as inflection points are added. For example, it is easier for people to understand that the robot has no preference by looking at Fig. 7e than Fig. 7g. The first part of H3 is supported.

Optimality ratio is significantly decreased from 0 to 4 compromise points ($F(1, 89) = 62.649, p < 0.001$) (Fig. 9e), indicating that path entropy reduces people's ability to identify the robot's true preference. For example, it is easier for people to understand that the robot has no preference from Fig. 7e than Fig. 7f. The first part of H4 is supported.

There is a significant interaction between the numbers of inflection and compromise points on optimality ratio, $F(2, 178) = 12.652, p < 0.001$. When the number of compromise points is high, the optimality ratio is significantly decreased from 2 to 4 inflection points ($p < 0.001$), while when number of compromise points is low, there is no significant difference ($p = 0.883$) (Fig. 9f). This indicates that when there are many compromise points, more inflection points exacerbates the detrimental effect of compromise points on optimality ratio, while when the number of compromise points is low, the detrimental effect is gone.

Optimality ratio is significantly higher when inflection points have the "same" configuration than when they have a "different" configuration ($F(1, 89) = 12.793, p = 0.001$) (Fig. 9g). This indicates that for maps without a preference, inflection points that move to the same type of terrain better reveal the robot's true (lack of) preference. For example, it is easier for people to understand that the robot has no preference by looking at Fig. 7h than Fig. 7g. The first part of H5 is supported. No other significant results were found.

Preference Range. We use a three-way repeated measures ANOVA to determine the effect of the number of inflection points, compromise points, and inflection point configuration on preference range (Table **??**). The number of inflection points has a significant effect on preference range ($F(2, 178) = 67.728, p < 0.001$). Post hoc analysis with a Bonferroni adjustment reveals that preference range is significantly decreased from 0 to 2 ($p < 0.001$) and from 0 to 4 ($p < 0.001$), but not from 2 to 4 inflection points ($p = 0.069$) (Fig. 10d). The second part of H3 is supported. Preference range is also significantly decreased from 0 to 4 compromise points ($F(1, 89) = 181.118, p < 0.001$) (Fig. 10e). The second part of H4 is supported.

There is a significant interaction between the numbers of inflection points and compromise points on preference range ($F(2, 178) = 18.848, p < 0.001$). When there are 4 compromise points, preference range is significantly decreased from 2 to 4 inflection points ($p = 0.003$), while when there are 0 compromise points, there is no significant difference ($p = 0.611$) (Fig. 10f). This indicates that inflection points have a detrimental effect on preference range only when they are exacerbated by compromise points, but that without the compromise points there is no detrimental effect.

Preference range was significantly decreased from "same" to "different" inflection point configuration ($F(1, 89) = 13.802, p < 0.001$) (Fig. 10g). This indicates that for maps without a preference, the preference range is lower when all inflection points have the "different" configuration than when the same number of inflection points have the "same" configuration. The second part of H5 is supported. No other significant differences are found.

Subjective Confidence. To determine the effect of inflection points, compromise

points, and inflection point configurations on subjective confidence, we conducted a generalized ordinal logistic model and estimated the model parameters through a generalized estimating equation (GEE) with AR(1) covariance structure (Table **??**). There is no significant effect of inflection points on subjective confidence (Fig. 11d). People are not significantly less confident about inferring the robot reasoning when dealing with demonstrations with more inflection points. The third part of H3 is rejected. Subjective confidence is significantly decreased from 0 to 4 compromise points ($p < 0.001$) (Fig. 11e). People are less confident about the robot's reasoning when dealing with demonstrations with more compromise points. The third part of H4 is supported. There were no significant effects of inflection point configuration on subjective confidence (Fig. 11g). The third part of H5 is rejected.

4.8 Discussion

People derive expectations about robot behavior by observing robot trajectories. Our work serves as a basis for enabling robots to use the trajectories they take to convey information about their reward functions. In this work, we introduce the concept of critical points and give two examples – inflection points and compromise points. Using these, we develop a method for systematically generating trajectories that possess the critical points we specify. We then test how trajectories with varying combinations of critical points affect human understanding of robot reward functions. We show that inflection points can have different effects on human understanding depending on whether a robot's reward function has particular terrain feature preferences or not. Specifically, when there is a preference for terrain features, adding inflection points improves human understanding, while when there is no preference, adding inflection points hinders understanding. In both cases, increasing the number of compromise points decreases human understanding of the robot's preferences.

Interestingly, our results showed that the subjective confidence did not increase with fewer compromise points as we expected. Future work is needed to understand why this is the case. For example, it is possible that if participants never saw the robot navigate over a rock, they would not be confident about what would happen if it *had* to navigate over a rock.

Additionally, our results showed that there was a significant effect of one pair of inflection points but no benefit to the second pair of inflection points suggesting that there is a "law of diminishing returns" in information conveyed by inflections. Because we only investigated two terrain types, one pair of inflection points is all that is necessary to indicate which terrain type is preferred. More work is needed to investigate whether our finding holds for more complex environments. For example, while we believe that one inflection point is needed to show relative preference between pairs of features, it is unclear whether the complexity of the path will overwhelm an observer rather than help them.

Finally, our study was performed in an online study and not on a real robot. We acknowledge that it may be difficult to modify real environments in order for optimal trajectories to include critical points. In environments where a real robot cannot demonstrate its reward function by adding inflection points, for example, it may be possible for the robot to display a simulated environment with a trajectory (such as those

we generated) to efficiently teach an observer about its preferences. Another option may be to demonstrate a non-optimal path that has more critical points. Future work is needed to understand whether our findings translate to real robots in real environments, and also whether other methods of demonstration are effective.

5 Conclusion & Future Work

This work serves as an example of how we can apply best practices sourced directly from humans in the design of robot explanation strategies. We summarize our findings in two different modes of explanation. We observe that the best examples of natural language goal specification take care to reduce the cognitive load the addressee must incur during comprehension. We also observe that when robots demonstrate their underlying objective function through their actions, they can be more interpretable if they ensure the strategic use of critical points.

The follow-up work is to complete the implementation of an explanation generation system on a robotic platform and test its performance in a physically deployed user-study. From here, there are two implementation-related areas to explore. First, there is the requirement of versatile scene representation (e.g. how do we represent the environment internally and allow it to be expressed via grounded language in different contexts). There is also the question of how to incorporate the human directly into the process. This development might be similar to designing a dialogue system. We will also need good methods of evaluating human-understanding in real-time. Incorporating a form of inverse-reinforcement learning for this purpose seems promising.

There are many potential avenues of further investigation in the theory of explanation strategies as well. One area is determining how to pick an appropriate strategy given a scenario. This would require studying how different strategies work across settings and tasks. For example, if a goal is not visible within the current scene, it might be more effective to communicate about the robot's objective function rather than the goal it is trying to reach.

6 Acknowledgements

The author would like to thank Dr. Stephanie Rosenthal, Prof. Siddhartha Srinivasa, and Dr. Henny Admoni for their patience and guidance throughout the exploration of this research topic. The author also thanks Dr. Jean Oh, Stefanos Nikolaidis, Shen Li, and the members of the Personal Robotics Lab at CMU.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. [Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution. Carnegie Mellon is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University. DM-0003432.

This work was (partially) funded by the DARPA SIMPLEX program through ARO contract number 67904LSDRP, National Institute of Health R01 (#R01EB019335), National Science Foundation CPS (#1544797), the Office of Naval Research, and the Richard K. Mellon Foundation.

References

[1] N. D. Allen, J. R. Templon, P. S. McNally, L. Birnbaum, and K. J. Hammond, "Statsmonkey: A data-driven sports narrative writer." in *Proc. AAAI Fall Symposium: Computational Models of Narrative*, 2010.

[2] P. Althaus, H. Ishiguro, T. Kanda, T. Miyashita, and H. I. Christensen, "Navigation for human-robot interaction tasks," in *Proc. ICRA*, vol. 2. IEEE, 2004, pp. 1894–1900.

[3] D. Avrahami-Zilberbrand and G. A. Kaminka, "Fast and complete symbolic plan recognition." in *Proc. IJCAI*, 2005, pp. 653–658.

[4] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[5] V. Bellotti and K. Edwards, "Intelligibility and accountability: Human considerations in context-aware systems," *Hum.-Comput. Interact.*, vol. 16, no. 2, pp. 193–212, Dec. 2001. [Online]. Available: http://dx.doi.org/10.1207/S15327051HCI16234_05

[6] R. Belvin, R. Burns, and C. Hein, "Development of the hrl route navigation dialogue system," in *Proc. International Conference on Human Language Technology Research*. Association for Computational Linguistics, 2001, pp. 1–5.

[7] A. Bicchi and G. Tonietti, "Fast and" soft-arm" tactics [robot arm design]," *IEEE Robotics & Automation Magazine*, vol. 11, no. 2, pp. 22–33, 2004.

[8] Y. Bisk, D. Marcu, and W. Wong, "Towards a dataset for human computer communication via grounded language acquisition," in *Proc. AAAI Workshop on Symbiotic Cognitive Systems*, 2016.

[9] S. N. Blisard and M. Skubic, "Modeling spatial referencing language for human-robot interaction," in *Proc. IEEE International Workshop on Robot and Human Interactive Communication*, 2005, pp. 698–703.

[10] D. Bryce, J. Benton, and M. W. Boldt, "Maintaining evolving domain models," in *Proc. IJCAI*. AAAI Press, 2016, pp. 3053–3059.

[11] H. H. Bui, "A general model for online probabilistic plan recognition," in *Proc. IJCAI*, vol. 3, 2003, pp. 1309–1315.

[12] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *Proc. International Conference on Healthcare Informatics (ICHI)*. IEEE, 2015, pp. 160–169.

[13] L. Carlson, M. Skubic, J. Miller, Z. Huo, and T. Alexenko, "Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task," *Topics in Cognitive Science*, vol. 6, no. 3, pp. 513–533, 2014.

[14] T. Chakraborti, G. Briggs, K. Talamadupula, Y. Zhang, M. Scheutz, D. Smith, and S. Kambhampati, "Planning for serendipity," in *Proc. IROS*. IEEE, 2015, pp. 5300–5306.

[15] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Explanation generation as model reconciliation in multi-model planning," *arXiv preprint arXiv:1701.08317*, 2017.

[16] T. Chakraborti, Y. Zhang, D. E. Smith, and S. Kambhampati, "Planning with resource conflicts in human-robot cohabitation," in *Proc. AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 1069–1077.

[17] R. Chatila, R. Alami, T. Simeon, J. Pettre, and L. Jaillet, "Safe, reliable and friendly interaction between humans and humanoids," in *Proc. IARP International Workshop on Humanoid and Human Friendly Robotics*, 2002, pp. 83–87.

[18] M. Cirillo, L. Karlsson, and A. Saffiotti, "Human-aware task planning for mobile robots," in *Proc. International Conference on Advanced Robotics*. IEEE, 2009, pp. 1–7.

[19] D. C. Dennett, *The intentional stance*. MIT press, 1989.

[20] A. K. Dey, "Explanations in context-aware systems." in *ExaCt*, 2009, pp. 84–93.

[21] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Proc. HRI*. ACM/IEEE, 2013, pp. 301–308.

[22] H. R. Everett, D. W. Gage, G. A. Gilbreath, R. T. Laird, and R. P. Smurlo, "Real-world issues in warehouse navigation," in *Photonics for Industrial Applications*. International Society for Optics and Photonics, 1995, pp. 249–259.

[23] G. Farin, "Class a bezier curves," *Computer Aided Geometric Design*, vol. 23, no. 7, pp. 573–581, 2006.

[24] ——, *Curves and surfaces for computer-aided geometric design: a practical guide*. Elsevier, 2014.

[25] K. Fischer and R. Moratz, "From communicative strategies to cognitive modelling," in *Proc. Workshop Epigenetic Robotics*, 2001.

[26] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 143–166, 2003.

[27] C. W. Geib and R. P. Goldman, "A probabilistic plan recognition algorithm based on plan tree grammars," *Artificial Intelligence*, vol. 173, no. 11, pp. 1101–1132, 2009.

[28] G. Gergely, Z. Nádasdy, G. Csibra, and S. Bíró, "Taking the intentional stance at 12 months of age," *Cognition*, vol. 56, no. 2, pp. 165–193, 1995.

[29] M. Göbelbecker, T. Keller, P. Eyerich, M. Brenner, and B. Nebel, "Coming up with good excuses: What to do when no plan can be found." *Cognitive Robotics*, no. 10081, 2010.

[30] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, *et al.*, "Designing robots for long-term social interaction," in *Proc. IROS*. IEEE, 2005, pp. 1338–1343.

[31] A. Herzig, V. Menezes, L. N. de Barros, and R. Wassermann, "On the revision of planning tasks," in *Proc. European Conference on Artificial Intelligence*. IOS Press, 2014, pp. 435–440.

[32] G. Hoffman and C. Breazeal, "Cost-based anticipatory action selection for human–robot fluency," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 952–961, 2007.

[33] ——, "Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team," in *Proc. HRI*. ACM, 2007, pp. 1–8.

[34] H. Horacek, "How to build explanations of automated proofs: A methodology and requirements on domain representations." in *Proc. ExaCt*, 2007, pp. 34–41.

[35] T. M. Howard, S. Tellex, and N. Roy, "A natural language planner interface for mobile manipulators," in *Proc. ICRA*. IEEE, 2014, pp. 6652–6659.

[36] M. J. Huber, E. H. Durfee, and M. P. Wellman, "The automated mapping of plans for plan recognition," in *Proc. International Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1994, pp. 344–351.

[37] K. Ikuta, H. Ishii, and M. Nokata, "Safety evaluation method of design and control for human-care robots," *The International Journal of Robotics Research*, vol. 22, no. 5, pp. 281–297, 2003.

[38] S. Ishii, S. Tanaka, and F. Hiramatsu, "Meal assistance robot for severely handicapped people," in *Proc. ICRA*, vol. 2. IEEE, 1995, pp. 1308–1313.

[39] K. Kamewari, M. Kato, T. Kanda, H. Ishiguro, and K. Hiraki, "Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion," *Cognitive Development*, vol. 20, no. 2, pp. 303–320, 2005.

[40] H. A. Kautz and J. F. Allen, "Generalized plan recognition." in *Proc. AAAI*, vol. 86, no. 3237, 1986, p. 5.

[41] O. Z. Khan, P. Poupart, and J. P. Black, "Minimal sufficient explanations for factored markov decision processes," in *Proc. ICAPS*, 2009.

[42] K. Kim and G. E. Fainekos, "Approximate solutions for the minimal revision problem of specification automata," in *Proc. IROS*. IEEE, 2012, pp. 265–271.

[43] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus, "Ikeabot: An autonomous multi-robot coordinated furniture assembly system," in *Proc. ICRA*. IEEE, 2013, pp. 855–862.

[44] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *Proc. HRI*. IEEE Press, 2010, pp. 259–266.

[45] K. M. Krishna, R. Alami, and T. Siméon, "Safe proactive plans and their execution," *Robotics and Autonomous Systems*, vol. 54, no. 3, pp. 244–255, 2006.

[46] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more?: the effects of mental model soundness on personalizing an intelligent agent," in *Proc. SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1–10.

[47] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *Proc. Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 2013, pp. 3–10.

[48] D. Kulić and E. A. Croft, "Real-time safety for human–robot interaction," *Robotics and Autonomous Systems*, vol. 54, no. 1, pp. 1–12, 2006.

[49] A. Kulkarni, T. Chakraborti, Y. Zha, S. G. Vadlamudi, Y. Zhang, and S. Kambhampati, "Explicable robot planning as minimizing distance from expected behavior," *arXiv preprint arXiv:1611.05497*, 2016.

[50] P. Langley, "Explainable agency in human-robot interaction," 2016.

[51] N. Lesh and O. Etzioni, "A sound and fast goal recognizer," in *Proc. IJCAI*, vol. 95, 1995, pp. 1704–1710.

[52] W. J. Levelt, "Perspective taking and ellipsis in spatial descriptions," *Language and Space*, pp. 77–107, 1996.

[53] S. C. Levinson, "Frames of reference and molyneuxs question: Crosslinguistic evidence," *Language and space*, pp. 109–169, 1996.

[54] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, "Learning and inferring transportation routines," *Artificial Intelligence*, vol. 171, no. 5-6, pp. 311–331, 2007.

[55] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proc. SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 2119–2128.

[56] L. J. Luotsinen, H. Fernlund, and L. Bölöni, "Automatic annotation of team actions in observations of embodied agents," in *Proc. International joint conference on Autonomous agents and multiagent systems*. ACM, 2007, p. 9.

[57] B. Lussier, A. Lampe, R. Chatila, J. Guiochet, F. Ingrand, M.-O. Killijian, and D. Powell, "Fault tolerance in autonomous systems: How and how much?" in *Proc. IARP-IEEE/RAS-EURON Joint Workshop on Technical Challenges for Dependable Robots in Human Environments (DRHE)*, 2005.

[58] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: Connecting language, knowledge, and action in route instructions," in *Proc. National Conference on Artificial Intelligence*, 2006.

[59] D. L. McGuinness and A. Borgida, "Explaining subsumption in description logics," in *Proc. IJCAI*, 1995, pp. 816–821.

[60] M. Molineaux, U. Kuter, and M. Klenk, "What just happened? explaining the past in planning and execution," DTIC Document, Tech. Rep., 2011.

[61] J. D. Moore and W. R. Swartout, "A reactive approach to explanation: taking the users feedback into account," in *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Springer, 1991, pp. 3–48.

[62] R. Moratz, K. Fischer, and T. Tenbrink, "Cognitive modeling of spatial reference for human-robot interaction," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 04, pp. 589–611, 2001.

[63] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa, "Game-theoretic modeling of human adaptation in human-robot collaboration," in *Proc. HRI*, 2017.

[64] M. Nokata, K. Ikuta, and H. Ishii, "Safety-optimizing method of human-care robot design and control," in *Proc. ICRA*, vol. 2. IEEE, 2002, pp. 1991–1996.

[65] S. Nonaka, K. Inoue, T. Arai, and Y. Mae, "Evaluation of human sense of security for coexisting robots using virtual reality. 1st report: evaluation of pick and place motion of humanoid robots," in *Proc. ICRA*, vol. 3, April 2004, pp. 2770–2775 Vol.3.

[66] E. Pacchierotti, H. Christensen, and P. Jensfelt, "Embodied social interaction for service robots in hallway environments," in *Field and Service Robotics*. Springer, 2006, pp. 293–304.

[67] E. Pacchierotti, H. I. Christensen, and P. Jensfelt, "Design of an office-guide robot for social interaction studies," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 4965–4970.

[68] V. Perera, S. P. Selveraj, S. Rosenthal, and M. Veloso, "Dynamic generation and refinement of robot verbalization," in *Proc. RO-MAN*. IEEE, 2016, pp. 212–218.

[69] K. Perlin, "An image synthesizer," *Proc. SIGGRAPH Computer Graphics*, vol. 19, no. 3, pp. 287–296, 1985.

[70] ——, "Improving noise," in *Transactions on Graphics*, vol. 21, no. 3. ACM, 2002, pp. 681–682.

[71] K. Perlin and E. M. Hoffert, "Hypertexture," in *Proc. SIGGRAPH Computer Graphics*, vol. 23, no. 3. ACM, 1989, pp. 253–262.

[72] D. V. Pynadath and M. P. Wellman, "Generalized queries on probabilistic context-free grammars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 65–77, 1998.

[73] M. Ramırez and H. Geffner, "Plan recognition as planning," in *Proc. IJCAI*, 2009, pp. 1778–1783.

[74] M. Ramírez and H. Geffner, "Probabilistic plan recognition using off-the-shelf classical planners," in *Proc. AAAI*. AAAI Press, 2010, pp. 1121–1126.

[75] ——, "Goal recognition over pomdps: Inferring the intention of a pomdp agent," in *Proc. IJCAI*. AAAI Press, 2011, pp. 2009–2014.

[76] R. Rao, K. Conn, S.-H. Jung, J. Katupitiya, T. Kientz, V. Kumar, J. Ostrowski, S. Patel, and C. J. Taylor, "Human robot interaction: application to smart wheelchairs," in *Proc. ICRA*, vol. 4. IEEE, 2002, pp. 3583–3588.

[77] S. Rosenthal, S. P. Selvaraj, and M. Veloso, "Verbalization: Narration of autonomous robot experience," in *Proc. IJCAI*. AAAI Press, 2016, pp. 862–868.

[78] B. Seegebarth, F. Müller, B. Schattenberg, and S. Biundo, "Making hybrid plans more clear to human usersa formal approach for generating sound explanations," in *Proc. International Conference on International Conference on Automated Planning and Scheduling*. AAAI Press, 2012, pp. 225–233.

[79] J. Shah, J. Wiken, B. Williams, and C. Breazeal, "Improved human-robot team performance using chaski, a human-inspired plan execution system," in *Proc. HRI*. ACM, 2011, pp. 29–36.

[80] M. T. Shaikh, M. A. Goodrich, and D. Yi, "Adverb palette: Gui-based support for human interaction in multi-objective path-planning," in *Proc. HRI*. IEEE Press, 2016, pp. 515–516.

[81] M. T. Shaikh, M. A. Goodrich, D. Yi, and J. Hoehne, "Interactive multi-objective path planning through a palette-based user interface," in *Proc. SPIE Defense + Security*. International Society for Optics and Photonics, 2016, pp. 98 370K–98 370K.

[82] E. A. Sisbot, L. F. Marin-Urias, R. Alami, and T. Simeon, "A human aware mobile robot motion planner," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 874–883, 2007.

[83] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock, "Spatial language for human-robot dialogs," *Proc. SMC Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 154–167, 2004.

[84] S. Sohrabi, J. A. Baier, and S. A. McIlraith, "Diagnosis as planning revisited." in *Proc. International Conference on Principles of Knowledge Representation and Reasoning*, 2010.

[85] K. Talamadupula, G. Briggs, T. Chakraborti, M. Scheutz, and S. Kambhampati, "Coordination in human-robot teams using mental modeling and plan recognition," in *Proc. IROS*. IEEE, 2014, pp. 2957–2962.

[86] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proc. AAAI*, San Francisco, CA, August 2011, pp. 1507–1514.

[87] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *Proc. SMC Part A (Systems and Humans)*, vol. 35, no. 4, pp. 460–470, 2005.

[88] M. Veloso, N. Armstrong-Crews, S. Chernova, E. Crawford, C. McMillen, M. Roth, D. Vail, and S. Zickler, "A team of humanoid game commentators," *Proc. International Journal of Humanoid Robotics*, vol. 5, no. 03, pp. 457–480, 2008.

[89] D. Voelz, E. André, G. Herzog, and T. Rist, "Rocco: A robocup soccer commentator system," *RoboCup-98: Robot Soccer World Cup II*, pp. 50–60, 1999.

[90] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic bulletin & review*, vol. 1, no. 2, pp. 202–238, 1994.

[91] D. Yi, M. A. Goodrich, and K. D. Seppi, "Informative path planning with a human path constraint," in *Proc. SMC*. IEEE, 2014, pp. 1752–1758.

[92] D. Yi and M. A. Goodrich, "Supporting task-oriented collaboration in human-robot teams using semantic-based path planning," in *Proc. SPIE*, vol. 9084, 2014.

[93] D. Yi, M. A. Goodrich, and K. D. Seppi, "Homotopy-aware rrt*: Toward human-robot topological path-planning," in *Proc. HRI*. IEEE, 2016, pp. 279–286.

[94] D. Yi, T. M. Howard, M. A. Goodrich, and K. D. Seppi, "Expressing homotopic requirements for mobile robot navigation through natural language instructions," in *Proc. IROS*. IEEE, 2016, pp. 1462–1468.

[95] Y. Zhang, H. H. Zhuo, and S. Kambhampati, "Plan explainability and predictability for cobots," *CoRR*, vol. abs/1511.08158, 2015.

[96] M. Zinn, O. Khatib, B. Roth, and J. K. Salisbury, "Playing it safe [human-friendly robots]," *IEEE Robotics & Automation Magazine*, vol. 11, no. 2, pp. 12–21, 2004.