The Effectiveness of Dynamically Processed Incremental Descriptions in Human Robot Interaction

CHRISTOPHER D. WALLBRIDGE, IROHMS, University of Cardiff and University of Plymouth, UK
ALEX SMITH, MANUEL GIULIANI, and CHRIS MELHUISH, Bristol Robotics Laboratory, UK
TONY BELPAEME, IDlab - imec, Ghent University and University of Plymouth, Belgium
SÉVERIN LEMAIGNAN, Bristol Robotics Laboratory, UK

We explore the effectiveness of a dynamically processed incremental referring description system using underspecified ambiguous descriptions that are then built upon using linguistic repair statements, which we refer to as a dynamic system. We build a dynamically processed incremental referring description generation system that is able to provide contextual navigational statements to describe an object in a potential real-world situation of nuclear waste sorting and maintenance. In a study of 31 participants, we test the dynamic system in a case where a user is remote operating a robot to sort nuclear waste, with the robot assisting them in identifying the correct barrels to be removed. We compare these against a static non-ambiguous description given in the same scenario. As well as looking at efficiency with time and distance measurements, we also look at user preference. Results show that our dynamic system was a much more efficient method—taking only 62% of the time on average—for finding the correct barrel. Participants also favoured our dynamic system.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Cognitive robotics**; **Spatial and physical reasoning**; • **Human-centered computing** → *Collaborative interaction;*

Additional Key Words and Phrases: Human robot interaction, natural language, spatial referring expressions, dynamic description, ambiguous, machine learning, user study, robots for nuclear environments

ACM Reference format:
Christopher D. Wallbridge, Alex Smith, Manuel Giuliani, Chris Melhuish, Tony Belpaeme, and Séverin Lemaignan. 2021. The Effectiveness of Dynamically Processed Incremental Descriptions in Human Robot Interaction. *Trans. Hum.-Robot Interact.* 11, 1, Article 7 (October 2021), 24 pages.
https://doi.org/10.1145/3481628

This work was supported by UK Engineering and Physical Sciences Research Council (EPSRC No. EP/R02572X/1) for the National Centre for Nuclear Robotics (NCNR). This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. This work was also conducted with support of the Centre for Artificial Intelligence, Robotics and Human-Machine Systems (IROHMS) operation C82092 and partially funded by the European Regional Development Fund (ERDF) through the Welsh Government.
Authors' addresses: C. D. Wallbridge, IROHMS, University of Cardiff and University of Plymouth, Cardiff University, Cardiff, Wales, UK, CF10 3AT; email: wallbridgec@cardiff.ac.uk; A. Smith, M. Giuliani, C. Melhuish, and S. Lemaignan, Bristol Robotics Laboratory, T Block, University of the West of England, Frenchay, Coldharbour Ln, Bristol, UK; emails: {alex.smith, manuel.giuliani, chris.melhuish, severin.lemaignan}@brl.ac.uk; T. Belpaeme, IDlab - imec, Ghent University and University of Plymouth, Technologiepark-Zwijnaarde, Ghent, Belgium; email: tony.belpaeme@ugent.be.

2573-9522/2021/10-ART7 $15.00
https://doi.org/10.1145/3481628

1 INTRODUCTION

Currently in the UK, the Nuclear Decommissioning Authority (NDA) is investing in research and development to reduce the temporal and financial cost of decommissioning the Sellafield site in the north of England [1]. This is currently estimated to take around 100 years at a cost of £162 billion [26]. The National Centre for Nuclear Robotics (NCNR) was created to tackle some of the main issues surrounding this mammoth task and brings together a diverse pool of expertise in robotics, including sensor fusion, artificial intelligence, control theory, intelligent grasping, resilient embedded systems and radiation hardening.

In many decommissioning environments, human entry is either undesirable or impossible. "Dirty" areas, which may contain radioactive dust, require the use of air-fed suits for human intervention, which create secondary waste after use and also reduce the capability of the workers. Therefore, there is a strong desire to remove the need for human intervention in these environments completely, wherever possible, by introducing robotic systems and the application of advanced telemanipulation (including shared autonomy, low cognitive load interfaces, safety, etc.). The complexity of the decommissioning environments require a high level of autonomous control, particularly in tasks involving complex manipulation, such as sorting a waste pile. Safety critical guidelines specified by the nuclear industry require a human in the loop; this means that robotic systems must work under shared control, which can be defined under different levels of autonomy. In addition, for many of the tasks in the decommissioning environments, a human is required for complex visual recognition or physical manipulation outside the scope of current robotics progress.

To enhance human robot collaboration, we look into generating spatial referring expressions to provide natural descriptions for people on the location of an object in such a hazardous environment.

We use spatial referring expressions to describe objects or locations by where they are—for example, "You left your keys under the folder on the desk" [27]. Spatial referring expressions are often used by people to identify an object, even in the case where another identifier—such as colour—could be used. Even where another indicator identifies an object, such as circling the object in an augmented reality system, people still like a description as well [30]. This can be useful for an assistive robot, such as in the previous example which assists you in locating your personal belongings.

The usual assumption in robotics is that a sentence uniquely describing the location of the object is best. In this article, we refer to such descriptions as "non-ambiguous", as they attempt to create a description that identifies a single location or object that will not allow it to be confused with another. This seems to stem from early work on referring expression generation such as the incremental algorithm [5]. Such systems cite Gricean maxims [10]:

- *Maxim of Quantity:* The referring expression should provide the relevant information without extra extraneous information.
- *Maxim of Quality:* The referring expression should be true.
- *Maxim of Relation:* The referring expression should be relevant.
- *Maxim of Manner:* The referring expression should be clear as to its contribution, be timely and avoid ambiguity.

Such non-ambiguous statements attempt to take a very rigid approach to the Maxim of Manner. Depending on the task, this can result in a combinatorial explosion if the problem space is large or complex as the possible number of descriptions increases at a possible rate of $N!$, where N is the number of objects that could be referred to. A lot of work focuses on trying to reduce this combination explosion, such as by looking at landmarks to narrow down the list of targets [12] or by only taking into account the objects known to the user [15, 23].

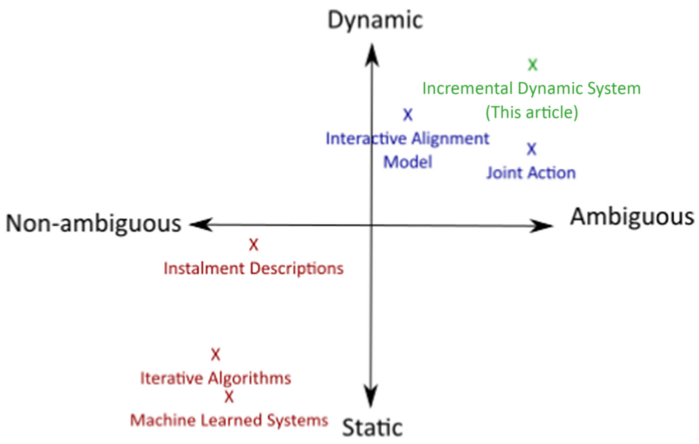

Fig. 1. A graph plotting current literature of referring expressions on the axes of non-ambiguous to ambiguous and static to dynamic. Much of the literature on computationally generating referring expressions (marked in red) is very static and emphasises attempts to be non-ambiguous. By contrast, many socio-linguistic models (marked in blue) show that interactions between people are much more dynamic and likely to be ambiguous without context. This work (marked in green) attempts to develop an interaction method much more like human-human interaction.

Although natural language generation will generate non-ambiguous descriptions, people often are ambiguous when describing spatial locations, requiring online disambiguation [24]. Between people, the disambiguation exchange is fast paced and dynamic, and therefore this process is difficult to replicate on robots. Disambiguation in robots requires several dialogue turns to narrow the referent, each turn often relying on planning [11] and confirmation, such as with instalment descriptions [9]. However there are many methods that can help disambiguate, such as relying on situational context [19], or realising that a single statement can imply multiple actions [25].

Attempts have been made to generate evaluation frameworks for spatial referring expressions. These often base their evaluation upon a single direction of communication [31]. As such, this often places a higher focus on the description itself rather than the fact that the description is meant to be part of a communicative act [13].

Where referring expression generation systems take into account two-way communication, the focus still is on adding description until no ambiguity remains. DeVault et al. [6] based a system on the incremental algorithm that would keep adding information until the user confirmed they had found the correct target. Instalment description [9] develops upon this by breaking down complex descriptions into smaller episodes, requiring confirmation at each stage. However, these instalment descriptions still rely on calculating the description at the beginning of the interaction, and require confirmation of having followed that part of the instruction, rather than being able to change with an updating situation.

In socio-linguistics, describing the location of something is often a two-way communicative exchange [4]. In such situations, people will not use a full description, but rather an under-specified description, that they then correct with a dynamic strategy of linguistic repair. This can help to share the cognitive load between the describer and the person who is listening to the description. This method also means that each participant contributes to the description until a shared grounding criterion is met [3]. It is only when people are struggling to reach such an alignment that a full description becomes necessary [22], although this may also be dependent on the task that agents are attempting to collaborate on [7] (see Figure 1 for a comparison of the literature). This process

can be highly dynamic, especially as seen in children, where a child receiving a description may take actions to prompt the describer to repair or allow for a simpler description [29]. In interactions between children, four types of spatial referring expression are identified:

- *Ambiguous-descriptive:* These statements provide details on the position of the object but refer to more than one possible location.
- *Contextual:* These are statements that follow on from a previous statement or action and would make no sense to a third party who entered the conversation at the time the statement was made (e.g., "The other one").
- *Negation:* This is when no statement is made other than to indicate the location chosen was incorrect (e.g., "No").
- *Non-ambiguous:* These statements can only refer to one possible location.

Previous research looked at using an incremental dynamic referring description using a simple under-specified statement to begin with, and then relying on a strategy of follow-up repair statements based on the resulting interaction—for a city planning game. This dynamic system was compared to non-ambiguous descriptions [28]. It was found that in an unfamiliar task, dynamic descriptions were more efficient in terms of time. In a second round of the game, this effect disappeared. However, it was also found that for this game-like task, there was a preference from participants for non-ambiguous descriptions. The reason often given was a preference for up-front information. There was a consideration that the task was very short, with each round taking approximately 90 seconds on average, and that extended interactions may prove non-ambiguous descriptions to eventually be better in longer interactions. In addition, although four types of description were identified in Wallbridge et al. [29], only two of them—ambiguous-descriptive and negation—were generated by the robot in this simple scenario.

In this article, we look at a possible real-use scenario, in which a remote piloted robot is used to sort nuclear waste. We develop a dynamically processed incremental referring description generation system (hereby referred to as the dynamic system) for use in this real-world scenario. This work builds upon and differs from Wallbridge et al. [28] in the following ways:

Realism: This study looks at a much more realistic setup. We use a robot that actually manipulates the world, in a potential real use-case scenario. This transfers from a much more toy like or abstract scenario into something that could be used in the real world, with the difficulties—such as sensor noise—that entail.

Continuous domain: The work presented here looks at a continuous world rather than a grid world.

Increased complexity of dynamic description: The dynamic description involves much more complexity, adding in localised non-ambiguous descriptions, and refining contextual descriptions.

Clear result: The dynamic incremental descriptions are much more efficient in all cases, unlike in previous work where there were questions of whether the non-ambiguous descriptions could be more effective given more time. This study involves a much longer scenario, and with users given more time to familiarise themselves with the equipment.

2 METHODOLOGY

2.1 Research Question

We wanted to investigate what type of description strategy would be more efficient and preferable to a person remote piloting a robot to locate objects, dynamic or non-ambiguous. These description strategies are defined as follows:

Fig. 2. The robot is remote piloted by a user to pick up radioactive barrels to a marked area for later disposal.

- *Dynamic:* An ambiguous description is used to initiate—for example, "A grey barrel is next to a yellow barrel". Based on the user's actions when piloting the robot, we then follow up with repair statements in an attempt to guide the user to the correct object—for example, "turn left" or "not that way".
- *Non-ambiguous:* A complete description is given of the object that is trying to be located, such that it cannot refer to any other object—for example, "A grey barrel is next to a yellow barrel and next to a grey barrel".

2.2 The Task

We developed a scenario in which certain nuclear waste, stored in barrels, needs to be sorted out of a storage area. At the time of this writing, there are nearly 1 million barrels of various types of nuclear waste being stored in facilities around the Sellafield site. These barrels need to be monitored and maintained regularly, a task which is currently performed by human operators but would preferably be automated as much as possible. A PAL Robotics TIAGo robot[1] was placed in a room with 50 model barrels of six different colours—8 Chrome, 10 Green, 10 Grey, 2 Rusty, 10 Silver and 10 Yellow (Figure 2). Some of these barrels were identified as causing problems and needed to be sorted for safe disposal. These barrels cannot be readily identified from the static CCTV cameras, therefore requiring an on-site robot to identify them. This is representative of real scenarios, where radiation-hot rooms have poor remote visibility and can only be internally inspected either with the use of air-fed-suited workers or remotely controlled robots.

Participants had a screen on which they could see the controls for the robot, the previous message the robot had sent and other status information for the robot (Figure 3). Participants were able to steer the robot using keyboard arrow keys. They were able to execute a grab action using the "g" key. Finally, using the "f" key switched between a fast driving mode and a slow driving mode to make it easier to line up the robot on the barrels. They also had a live stream from the robot's camera head camera; this was in a fixed position that also showed a target circle, allowing them to see where the robot's arm would grab.

[1] https://tiago.pal-robotics.com/.

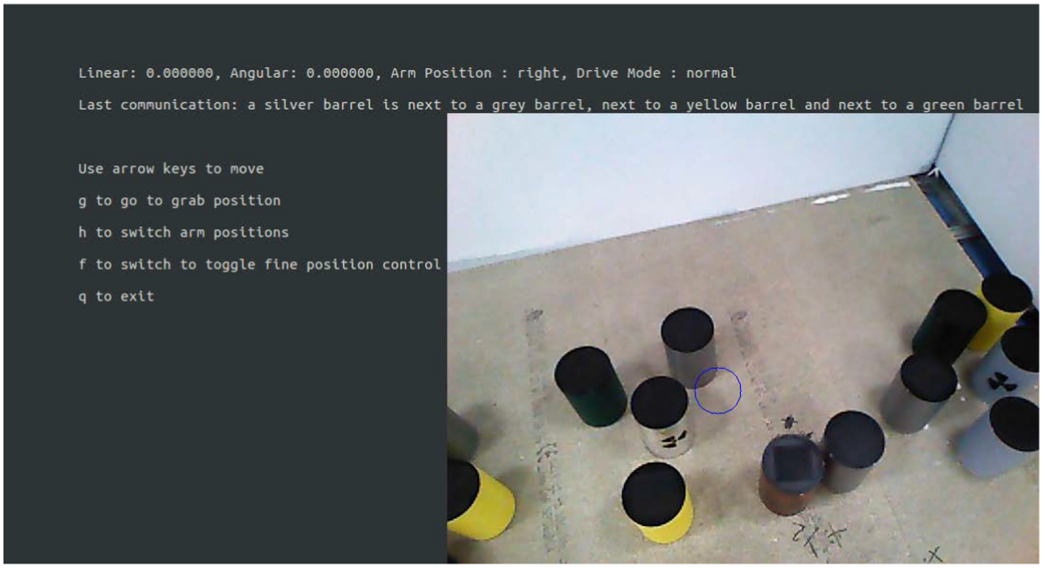

Fig. 3. The participant's screen during the task. The controls were displayed as well as information about the state of the robot, including the most recent communication sent. A stream of the robot's head camera, as well as a targeting reticule (the blue circle), allowed participants to accurately align the robot's position to grab a barrel.

Participants also had a view from four security cameras set up in the room (Figure 4). Picking up the barrels was accomplished with the use of an electromagnet. The electromagnet was attached to the end of the robot's arm and could be controlled by a remote switch that was given to the participants.

The task was divided into two rounds, with six target barrels in each round. The six barrels removed in the first round were not replaced for the second round, so the second round started with 44 barrels. The target barrels were always presented in the same order. Objects were designed to require between two and four spatial descriptors to describe non-ambiguously—for example, a non-ambiguous statement requiring two descriptors would be as follows:

> "A yellow barrel is next to a rusty barrel and next to a silver barrel".

In each map, one target required four descriptors, two required three descriptors and three required two descriptors.

Once the barrel was picked up, it had to be brought back to a marked area for later disposal. Timing is taken from the moment the robot starts its first description of the barrel and until a successful grab on the correct target is initiated. Participants would be informed if they were grabbing the wrong target upon initiating a grab on it.

2.3 Hypotheses

We designed our study to test the following hypotheses:

*H*1: A robot giving a dynamic description (an ambiguous initial statement with follow-up repair) would allow a person to find the correct barrel faster than when a robot is giving a non-ambiguous description.

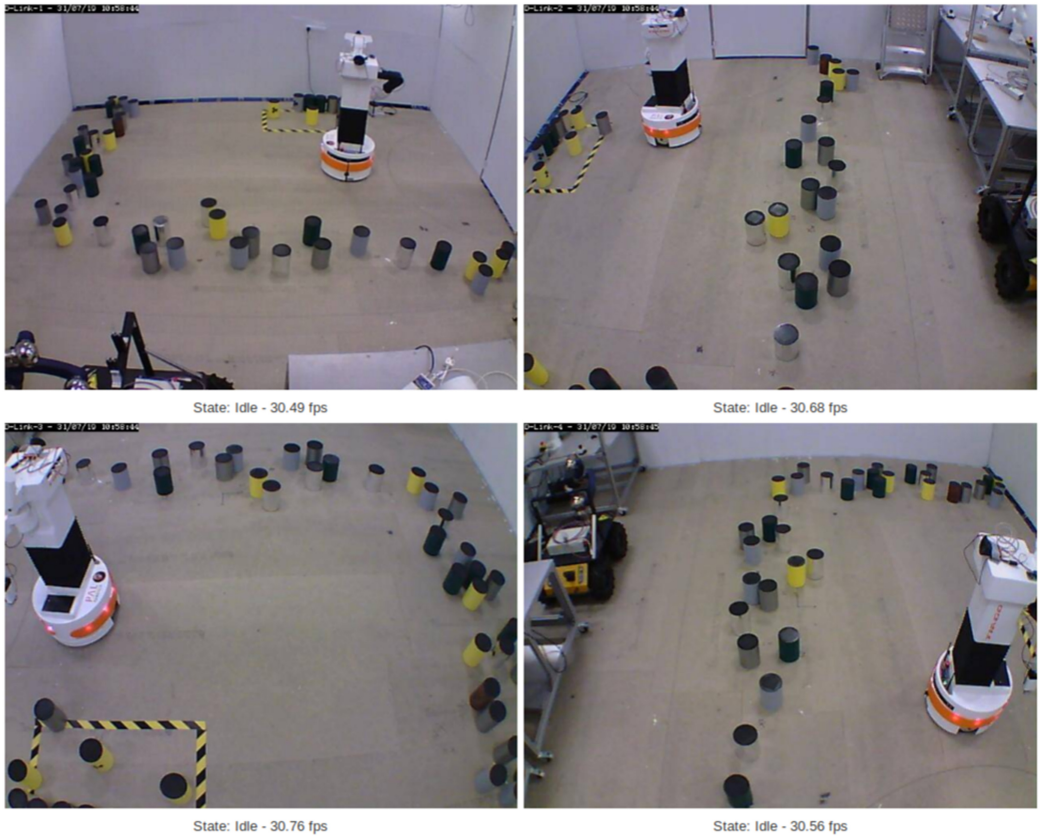

Fig. 4. The four static camera views available to the participant and the experimenter of the inside of the room. Barrels, some of which the participants were told were emitting radiation, were placed around a room. Participants had to collect these barrels and bring them back to a marked zone for later disposal.

H2: A robot giving a dynamic description would be prefered by the operator over a robot giving a non-ambiguous description.

2.4 Implementation

In addition to the TIAGo robot and the four cameras inside the room, we also used the HTC Vive base stations and a beacon. These were used to provide the coordinates of the barrels and the robot to the system. Using Underworlds [16], we represent the state of the world using 3D meshes. This allows the system to reason about the relations between objects based on their location. Spatial relations are calculated using bounding boxes for identifying nearby barrels as landmarks [12] to perform simple geometric comparisons. From these spatial affordances, we build a natural language description. For this study, the initial positions and meshes were provided to Underworlds, as well as the labels for objects, as acquiring these was not the focus of this work.

2.4.1 Non-Ambiguous System. In the case of generating a non-ambiguous description, in the same way as the incremental algorithm [5], we add descriptors in a greedy fashion that remove ambiguity until none remains. These descriptors are based on the colours of the barrels around our target. Due to the fact that we have five different camera angles which the participant could

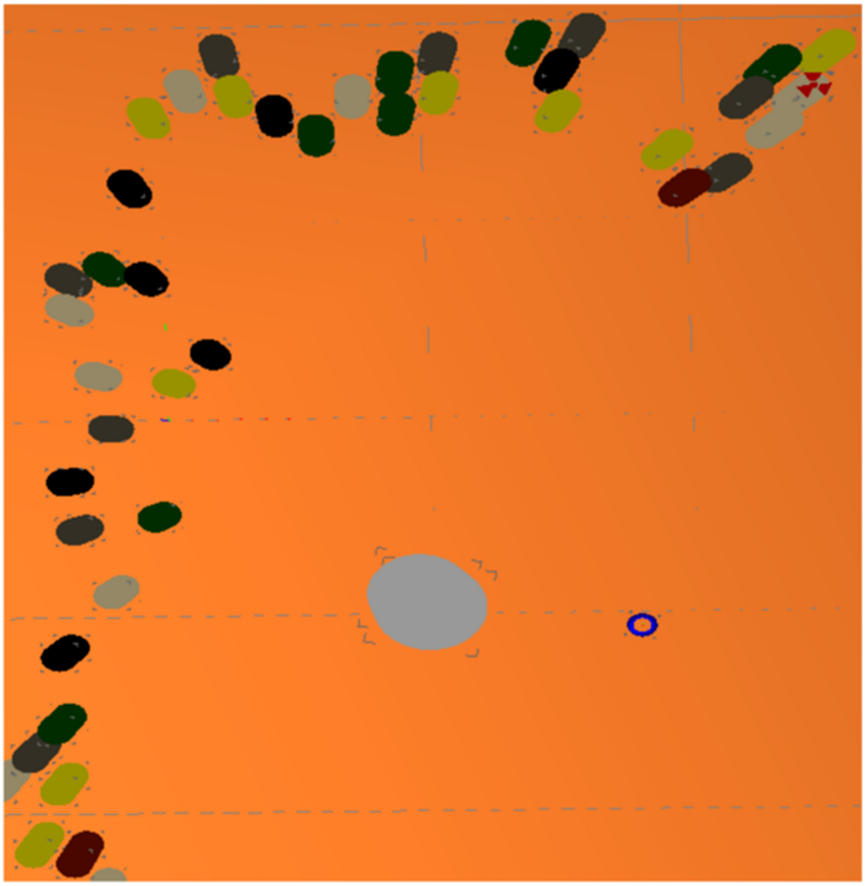

Fig. 5. The representation of the state of the world in Underworlds. The current target barrel is marked with a radiation symbol. The large cylinder represents the location of the robot, with the blue ring representing the approximate location of the robot's arm if it were to enter the grab position.

view at any one time—and we are not tracking where the participant is looking—we had to rely on "next to", rather than directions such as "to the left" or "right", to describe the positions of the barrels. Descriptions were repeated every 20 seconds after the robot had finished speaking, until the barrel was successfully picked up. An example of a non-ambiguous description requiring three descriptors would be as follows:

"A grey barrel is next to a grey barrel, next to a silver barrel and next to a green barrel".

2.4.2 Dynamically Processed Incremental Referring Description Generation System. As in previous work [28], we built the dynamic system to be fully autonomous. We conducted a pre-study (n = 14) in which we had two participants interacting, one acting as the robot's pilot as normal and one who acted as a describer. The describing participant could see the Underworlds representation of the world (Figure 5). In the Underworlds representation, they could see the target barrel marked with a radiation symbol and would have to guide the pilot to the correct barrel. After completing the first map, the participants switched roles. From these interactions, we sample the position the robot's arm would be if in the grab position at a rate of 20 Hz. Taking these coordinates, we

Table 1. Some Examples of the Training Data, Representing the State of the Interaction, and the Expected Output from the Classifier

Dist. to Targ.	C. Dist. to Targ.	Mag. of Motion	C. Yaw	Yaw Req.	Req. C. Yaw	Output
1.544663201	0.019257559	0.025458991	0.033564351	2.090630074	−0.520064175	0
1.659268017	−0.00011127	0.00017128	4.00E-05	−2.790334444	0.884704126	1
1.906600371	−0.016265772	0.020631222	0.008862872	2.088658345	−0.200643514	2
0.051776	0.000228	0.000313	−6.27E-05	2.922887	−0.45289	3

Dist.: Distance; C.: Change in; Mag.: Magnitude; Req.: Required.

calculated the distance to target, change in distance to target, magnitude of motion from the previous position, yaw required, change in yaw and required change in yaw to represent the state of the interaction. This gave us a total of 159,384 data points representing the state. Using recordings of the interactions, we then assign each of these states a classification—represented by a number in the data—in one of the following categories, similar to the previous study:

- *0—Negate*: A negative response indicating that the manipulator is heading in the wrong direction (e.g., "stop").
- *1—Contextual*: These are statements that follow on from a previous statement or action and would make no sense to a third party who entered the conversation at the time the statement was made. For the most part, in this scenario these were navigational statements (e.g., "Turn Left").
- *2—Positive*: A positive response given to the manipulator to indicate they are heading in the right direction (e.g., "Yes").
- *3—Localised Non-ambiguous*: Once the robot was close enough to the barrel, often a description would be used to uniquely describe the barrel compared to the ones around it. This was partly as sensor noise meant the describer could not be exactly sure which barrel the pilot was actually looking at (e.g., "The green barrel").

An example of our training data can be seen in Table 1.

For the classifier used in the dynamic condition in this study, we trained a **Multilayer Perceptron (MLP)** network.[2] This network used three fully connected hidden layers, each of size 20, a ReLU activation and a Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) solver. The decision to use these was based on the previous work in Wallbridge et al. [28]. We used 127,304 of the classified data points as a training set for the network. The remaining 32,080 records were used to test against. Testing the trained network against this testing set, we obtained a 98.5% success rate. The confusion matrix can be seen in Table 2. Although this is a high success rate for such a task, a large portion of the data—such as when the robot is idle—could be filtered with a simple algorithm. However, we still need the machine learning for more complex behaviours.

We also performed a second pre-study ($n = 4$) in which we gathered feedback on the dynamic system that allowed us to make some adjustments to the dynamic descriptions, such as including approximate angles on large turns.

For the dynamic description, we used a manually programmed timing mechanism due to time constraints. The timings were developed based on best judgement and iteration. The feedback was based on the average result from the classifier over the previous half a second—sampled at 20 Hz: if a negate statement was required, the robot would respond immediately. Positive statements were only given if the robot had not spoken in the previous 2 seconds. A contextual or localised non-ambiguous statement would be given if the robot had not spoken for 1 second if the information

[2]This network was trained using the scikit-learn library for Python [21].

Table 2. Confusion Matrix of the MLP Classifier Used for the Dynamic Condition When It Was Used on Our Test Data

		Prediction			
		Negate	Contextual	Positive	Localised Non-Ambiguous
Actual	Negate	1,497	146	1	5
	Contextual	57	13,823	92	16
	Positive	2	123	2,513	15
	Localised Non-Ambiguous	6	13	3	13,768

The data used to train and test the MLP was obtained from a series of pilot studies.

was different to the previous. Otherwise, for a contextual statement, the robot would repeat the previous statement after 5 seconds, and 10 seconds for localised non-ambiguous descriptions.

The dynamic description starts by giving an initial ambiguous statement using a single descriptor. From there, the classifier provides further descriptions based on the user's actions. These descriptions were then turned into sound using pyttsx.[3] If a negate statement is required, it would randomly select between saying "stop", "hold on" and "not that way". Positive statements would randomly select between "yeah" and "keep going". If a contextual statement was required, then some calculations were made to choose what to say. If the required change in yaw is less than 0.2 radians, then the robot would say "go forward". Otherwise, the robot would say the appropriate turn left or turn right. If the change in yaw required was also above 2.749 radians, it would say "about 180 degrees"; if it was over 2.051, then "about 135 degrees"; and over 1.178, then about "about 90 degrees". The localised non-ambiguous statements were generated in the same way as the non-ambiguous statements, but only looking at barrels near the current target. In the following, we give an example of an interaction using the dynamic system:

TIAGo—Initial Description: "A silver barrel is next to a chrome barrel." (0:00–0:03 seconds)
TIAGo—Contextual Navigation: "Turn left about 90 degrees." (0:06–0:10)
Participant: <Starts turning the robot left> (0:09–0:12)
TIAGo—Positive: "Keep going." (0:12–0:14)
Participant: <Continues turning> (0:12–0:15)
TIAGo—Contextual Navigation: "Go forward." (0:15–0:17)
Participant: <Moves the robot forward> (0:16–0:19)
TIAGo—Negation: "Stop." (0:17–0:19)
Participant: <Stops> (0:20)
TIAGo—Contextual Navigation: "Turn right." (0:22–0:23)
Participant: <Turns right> (0:24–0:25)
TIAGo—Contextual Navigation: "Go forward." (0:25–0:26)
Participant: <Moves forward> (0:27)
TIAGo—Localised Non-Ambiguous: "The silver barrel next to the chrome barrel." (0:28–0:32)
Participant: <Participant lines up carefully and successfully selects the correct barrel> (0:32–0:48)

2.5 Questionnaire

To test our second hypothesis, we used a questionnaire. We built our questionnaire using elements of the Godspeed Questionnaire [2]. This questionnaire is designed to measure a user's perception

[3]Pyttsx is a Python package that provides cross-platform support for text-to-speech generation.

in a number of categories using questions with a 5-point Likert scale. We took the sections on anthropomorphism, likeability and intelligence from the Godspeed. All these questions were asked at three stages: before the task, after the first map and after completing the second map. In the initial part of the questionnaire, we also asked some basic questions on demographics: age, familiarity with robots and first language.

After each map, the questionnaire also had added questions on the robot's task performance. We did this using a Likert scale for both the timeliness, appropriateness of instructions and how helpful they found the robot overall. There was also a question of how much feedback the robot gave, from too little to too much.

In the final part of the questionnaire, we also had three open-ended questions. The first was on the difference between the behaviour of the robot on the first and second map. The second was on which behaviour they preferred and why. Finally, there was a section for other comments.

2.6 Interaction

We used a within-subject design. Participants would see both conditions (dynamic and non-ambiguous) across the two different maps, with the order they would see in being counter-balanced. The order the participants would see the conditions in was randomly determined before meeting the experimenter.

After signing the consent form, participants were asked to fill out the first part of the questionnaire. Before starting the main task, the experimenter gave participants a tutorial on controlling the robot. In this tutorial, the controls and the statuses they could see on their screen were explained. There was also an explanation of the barrels they would see within the game and the way the robot would describe them. They were also given a chance to practice picking up the barrels and bringing them back to the marked zone.

TIAGo itself also gave a brief explanation of the controls, as well as giving more information about the task, before starting the first description. The task then proceeded as described in Section 2.2 using one of the randomly assigned conditions.

After completing the first map, the participant was given the second part of the questionnaire. Participants were informed that "for the next round the task would remain the same as before, but the robot's behaviour has now changed". The task then proceeded with the condition that the participant had not yet seen.

Upon completion of the second map, participants were given the final part of the questionnaire. Participants were then debriefed on the purpose of the study. Typically, the entire interaction, including the completion of questionnaires, took 40 to 45 minutes. Participants were paid £10 worth of Amazon gift vouchers upon completion.

Data on the position of the robot and target was recorded while participants were attempting to find a barrel being described.

2.7 Secondary Study

Although the main study returned promising results, one can argue that the static non-ambiguous descriptions used in the study were sub-optimal, especially when considering that the dynamic descriptions are made from the perspective of the robot, whereas the static non-ambiguous descriptions were perspective-free. There is a concern that because of this the static non-ambiguous descriptions are intrinsically harder than the dynamic ones, which might provide an unintended advantage to the dynamic descriptions. To assess whether the perspective from which descriptions are made offers an advantage, we ran a second study in we compared robot-centric against perspective-free descriptions.

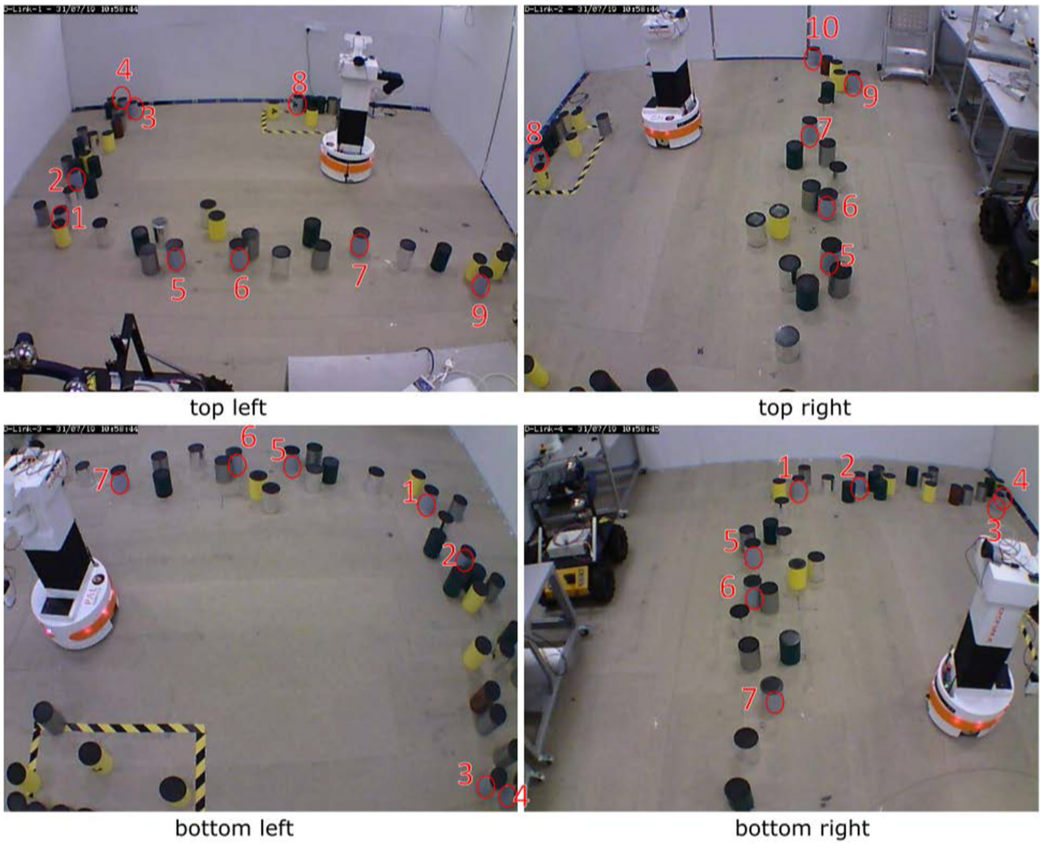

Fig. 6. An example of one of the five variants that was used in the secondary study looking at different types of static non-ambiguous descriptions. In this example, all the grey barrels have been labelled.

We took the picture from Figure 4 and created five variants of this picture, and in each one a different colour barrel would be labelled, to use in an online survey.[4] An example can be seen in Figure 6 with all the grey barrels labelled. In the study, we tested two different types of static non-ambiguous description. The first type used the same perspective-free descriptions we used in the main study, and the second type of descriptions were from the perspective of the robot. These were similar to the dynamic system descriptions from the main study but now contained a full non-ambiguous description which should allow the participant to identify the correct barrel from a single description. See Table 3 for all the descriptions that were used in this study.

Participants were given a tutorial before beginning the study. They were told to answer as quickly and accurately as possible. Each question was presented with the image with the barrels labelled at the top, in the middle they were given the description, and then they were asked to select the correct barrel. At the bottom, the key was shown again, to allow participants to refer to it. Once they submitted their answers, they were not able to change them. We measured participants on the time it took them to respond to questions, as well as gave them a score out of 10 based on how many answers were correct.

[4]You can find a version of this study for review and replication purposes here: https://cardiffunipsych.eu.qualtrics.com/jfe/form/SV_5C5nO5raHjMaTJQ.

Table 3. All the Descriptions Used in the Study to Assess the Influence of Perspective

No.	Target	Robot Perspective Descriptions	Perspective-Free Descriptions
1	Yellow 9	Turn right over 90 degrees, go forward, and pick up the yellow barrel that is next to the rusty barrel.	Pick the yellow barrel that is next to a rusty barrel and a yellow barrel.
2	Green 10	Turn right about 180 degrees, go forward, and pick up the green barrel next to a chrome barrel and a yellow barrel.	Pick the green barrel that is next to a yellow barrel and a chrome barrel.
3	Yellow 2	Turn left about 90 degrees, go forward, and pick up the yellow barrel that is next to a green barrel and another green barrel.	Pick the yellow barrel that is next to a green barrel, another green barrel and a silver barrel.
4	Grey 4	Turn left almost 90 degrees, go forward, and pick up the grey barrel that is next to a green barrel.	Pick the grey barrel that is next to a yellow barrel, a green barrel, a silver barrel and another grey barrel
5	Green 6	Turn left over 90 degrees, go forward and pick up the green barrel next to a chrome barrel.	Pick the green barrel that is next to a chrome barrel, a silver barrel and a grey barrel.
6	Grey 9	Turn right almost 180 degrees, go forward, and pick up the grey barrel next to a yellow barrel.	Pick the grey barrel that is next to a yellow barrel and next to another yellow barrel.
7	Green 1	Turn left 90 degrees, go forward and pick up the green barrel next to a chrome barrel.	Pick the green barrel that is next to a grey barrel, a chrome barrel and another green barrel.
8	Yellow 8	Turn right almost 180 degrees, go forward, and pick up the yellow barrel that is next to a grey barrel and a green barrel.	Pick the yellow barrel that is next to a yellow barrel, a grey barrel and a green barrel.
9	Green 3	Turn left about 90 degrees, go forward and pick the green barrel that is next to another green barrel and a silver barrel.	Pick the green barrel that is next to a silver barrel, a grey barrel, a yellow barrel and another green barrel.
10	Silver 4	Turn left almost 90 degrees, go forward and pick up the silver barrel that is next to a green barrel.	Pick the silver barrel that is next to a green barrel, a grey barrel, another grey barrel and a silver barrel.

The perspective-free descriptions were the same descriptions as used in the main study for our static non-ambiguous descriptions. The robot perspective descriptions were designed to be similar to the kind of descriptions the robot would use in the dynamic condition of our main study, but were formed to be a single non-ambiguous description.

We ran the study as a between-subject study on description type, with participants randomly assigned at the beginning of the study. Participants were recruited through Prolific.[5] Participants were restricted from Prolific to those with an approval rating of over 95%, fluent in English, were 18 or over, and were doing the study with a desktop or laptop. Participants were paid £1.88 for their time (a rate of £7.52 per hour for an estimated 15 minutes).

2.8 Demographics

2.8.1 Primary Study. We recruited 31 participants (8 F) from within the Bristol Robotics Laboratory. These were for the most part people with experience in or expertise on robots. This work was performed in the context of using collaborative robots to work with highly skilled nuclear decommissioning staff, who would have been trained to use the robots. Therefore, we consider using completely naive participants without training or knowledge in robotics as inappropriate. With a pool of more than 100 researchers within the Bristol Robotics Laboratory, we were able to recruit people who were naive to the purpose of the study. Participants had a mean age of 26.4 years (min = 17, max = 46 sd = 5.44). Sixteen of our participants did not speak English as a first language, but they were fluent enough as not to require being excluded from the study. Of the 31 participants, 15 (2 F) described themselves as roboticists, 6 (2 F) described themselves as having worked alongside robots and 8 (2 F) said they had interactions with commercial robots. Two (2 F) said they had not previous interacted with robots. Sixteen (3 F) of our participants saw the non-ambiguous condition followed by the dynamic. The remaining 15 (5 F) saw the dynamic condition followed by the non-ambiguous.

2.8.2 Secondary Study. We recruited 50 people to participate in the study through the Prolific crowdsourcing platform. One participant was removed, as it was evident based on the scores and time taken that they had not completed the study properly. This left us with 49 (23 F) participants in the study. As we were looking at just the descriptions, and participants would not be required to operate the robot, we believed it appropriate for any background to answer the survey. The mean

[5] https://www.prolific.co/.

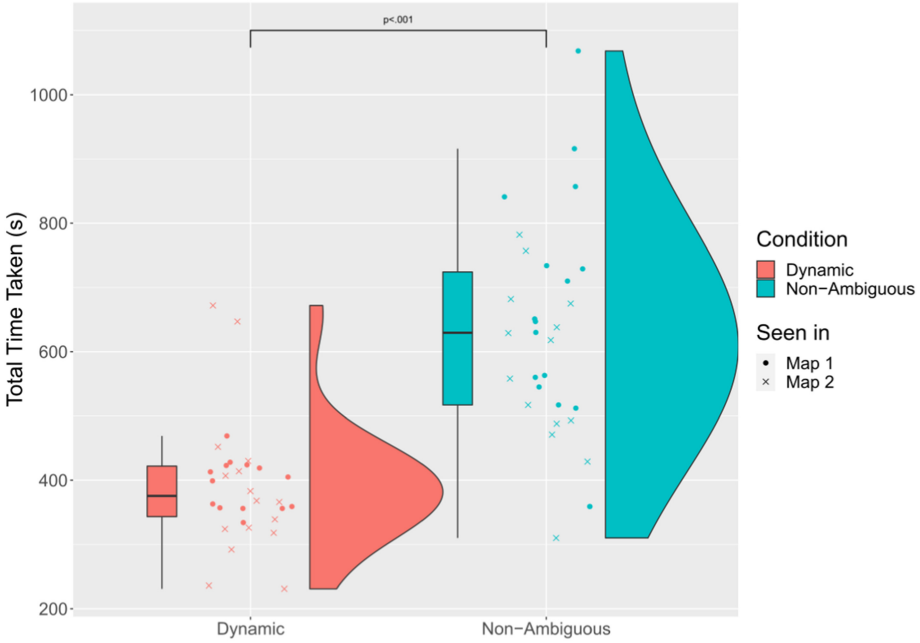

Fig. 7. The sum of time taken to correctly locate all barrels in a map. For each target barrel, the time is taken from the moment the robot starts the first description to the moment the successful grab was initiated. There is a significant difference between the time taken in the dynamic condition and non-ambiguous condition when both map 1 and map 2 are considered together.

age was 25.96 years (min = 18, max = 57, sd = 8.53). Twenty-three (14 F) participants saw the robot perspective description condition, with 26 (9 F) seeing the perspective-free description condition.

3 RESULTS

3.1 Primary Study: Task Performance

Due to a systems error, one participant did not log times for all parts of the study, so their data was not used in this analysis. No significant difference was found between maps (paired t-test: $t = 1.3583$, $df = 29$, $p = 0.185$, mean of map 1 = 541.29 seconds (sd = 192.57), mean of map 2 = 486.26 seconds (sd = 166.51)), as such we did not further consider order effects. We found a significant difference between the dynamic and non-ambiguous conditions (paired t-test: $t = -8.301$, $df = 29$, $p < 0.001$, mean of dynamic = 390.33 seconds (sd = 92.41), mean of non-ambiguous = 629.53 seconds (sd = 164.35)) (Figure 7).

As well as with the previous withdrawn participant, due to a sensor failure an additional participant did not log the distance travelled for all targets and thus was not used in this analysis. We also see a significant difference between the distances travelled across conditions (paired t-test: $t = -11.384$, $df = 28$, $p < 0.001$, mean of dynamic = 26.35 minutes (sd = 92.41), mean of non-ambiguous = 59.11 minutes (sd = 15.33)) (Figure 8).

3.2 Primary Study: Preference

For the analysis of the questionnaires, two participants were not considered due to failing to answer some of the questions. Figure 9 shows the overall combined scores for anthropomorphism, intelligence and likeability. Overall, in the category of anthropomorphism, we found a significant

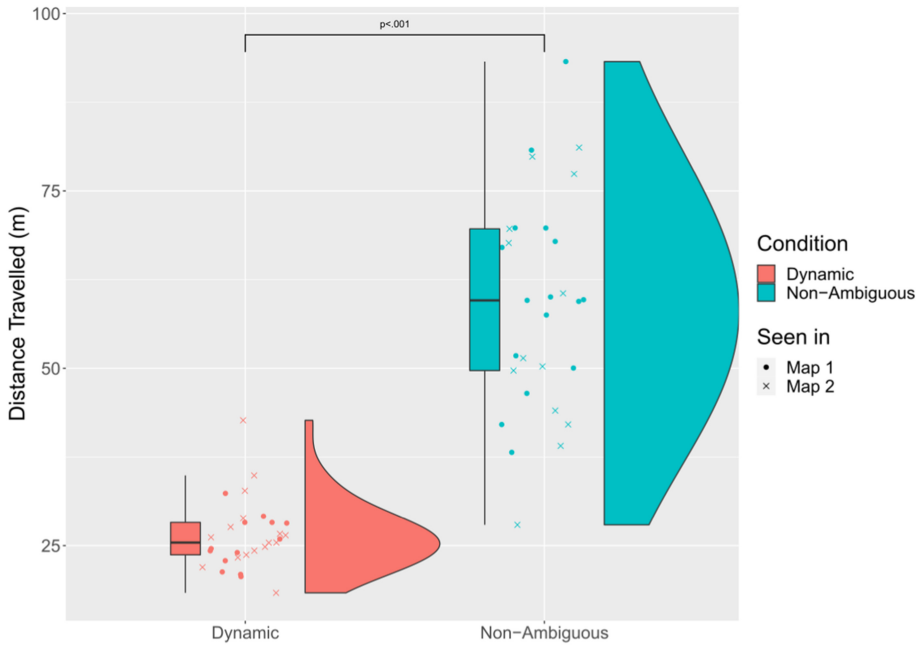

Fig. 8. The sum of distance travelled while correctly locating all barrels in a map. There is a significant difference between the distance travelled in the dynamic condition and non-ambiguous condition when both map 1 and map 2 are considered together.

difference between the dynamic and non-ambiguous conditions (paired t-test: $t = 4.3277$, $df = 28$, $p < 0.001$, mean of dynamic = 10.97 ($sd = 2.76$), mean of non-ambiguous = 8.48 ($sd = 2.77$)). Looking at the individual questions, we found a significant difference on "Artificial to Lifelike" between conditions (Wilcoxon signed-rank test with continuity correction: $V = 230$ $p = 0.018$, mean of dynamic = 2.31 ($sd = 1.04$), mean of non-ambiguous = 1.66 ($sd = 0.86$)). We also found a significant difference on the question of "Unconscious to Conscious" between conditions (Wilcoxon signed-rank test with continuity correction: $V = 259$, $p = 0.009$, mean of dynamic = 3.10 ($sd = 1.08$), mean of non-ambiguous = 2.17 ($sd = 0.89$). We did not see a significant difference between the dynamic and non-ambiguous conditions in the other two questions in the category of anthropomorphism (Figure 10).

In the overall category of likeability, we found a significant difference between the dynamic and non-ambiguous conditions (paired t-test: $t = 4.4384$, $df = 28$ $p < 0.001$, mean of dynamic = 15.24 ($sd = 2.91$), mean of non-ambiguous = 13.03 ($sd = 3.51$)). Looking at the individual questions, we found a significant difference on the question of "Dislike to Like" between conditions (Wilcoxon signed-rank test with continuity correction: $V = 241$, $p = 0.008$, mean of dynamic = 3.93 ($sd = 0.92$), mean of non-ambiguous = 3.10 ($sd = 1.29$)). We also found a significant difference on the question of "Unpleasant to Pleasant" between conditions (Wilcoxon signed-rank test with continuity correction: $V = 182$, $p = 0.018$, mean of dynamic = 3.76 ($sd = 0.83$), mean of non-ambiguous = 3.21 ($sd = 0.86$)). We did not see a significant difference between the dynamic and non-ambiguous in the other two questions on likeability (Figure 11).

In the category of intelligence, we found a significant difference between the dynamic and non-ambiguous conditions (paired t-test: $t = 4.9161$, $df = 28$, $p < 0.001$, mean of dynamic = 18.86 ($sd = 3.63$), mean of non-ambiguous = 15.69 ($sd = 4.05$)). Breaking down by question, we found a

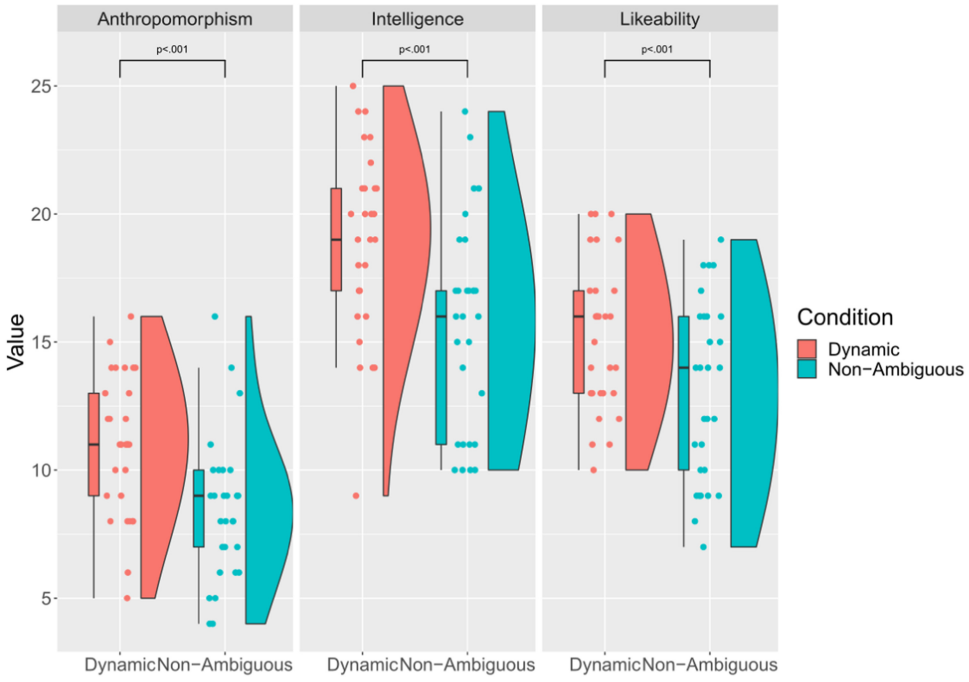

Fig. 9. Overall combined scores for the robot on questions of anthropomorphism, intelligence and likeability.

significant difference on "Ignorant to Knowledgeable" between conditions (Wilcoxon signed-rank test with continuity correction: $V = 212$, $p = 0.023$, mean of dynamic = 3.86 ($sd = 0.88$), mean of non-ambiguous = 3.03 ($sd = 1.02$)). We also found a significant difference on the question of "Incompetent to Competent" (Wilcoxon signed-rank test with continuity correction: $V = 171$, $p = 0.012$, mean of dynamic = 3.90 ($sd = 0.94$), mean of non-ambiguous = 3.07 ($sd = 1.13$)). In the question of "Unintelligent to Intelligent", a significant difference was found (Wilcoxon signed-rank test with continuity correction: $V = 239$, $p = 0.035$, mean of dynamic = 3.52 ($sd = 0.87$), mean of non-ambiguous = 2.90 ($sd = 0.90$)). We did not see a significant difference between the dynamic and non-ambiguous in the other two questions for intelligence (Figure 12).

A significant difference was found between dynamic and non-ambiguous when looking at the amount of feedback provided (Wilcoxon signed-rank test with continuity correction: $V = 190.5$, $p = 0.033$, mean of dynamic = 2.93 ($sd = 0.59$), mean of non-ambiguous = 2.41 ($sd = 1.02$)). On the question of "Inappropriate to Appropriate" feedback, we found a significant difference between our two conditions (Wilcoxon signed-rank test with continuity correction: $V = 234$, $p = 0.003$, mean of dynamic = 4.28 ($sd = 0.75$), mean of non-ambiguous = 3.28 ($sd = 1.31$)). "Unhelpful to Helpful" was also significant between dynamic and non-ambiguous (Wilcoxon signed-rank test with continuity correction: $V = 297$, $p < 0.001$, mean of dynamic = 4.41 ($sd = 0.68$), mean of non-ambiguous = 3.14 ($sd = 1.16$)). We did not see a significant difference on the remaining question of timeliness (Figure 13).

Of the 31 participants, 27 preferred the dynamic condition, with 3 preferring the non-ambiguous and 1 participant not preferring either condition. Of the three participants who did not prefer the dynamic condition, 1 participant stated he liked the challenge of finding the right barrel in the non-ambiguous condition. Another said that the dynamic condition "forced" them to do it its way, and the non-ambiguous condition allowed them to do the task in their own way. The final

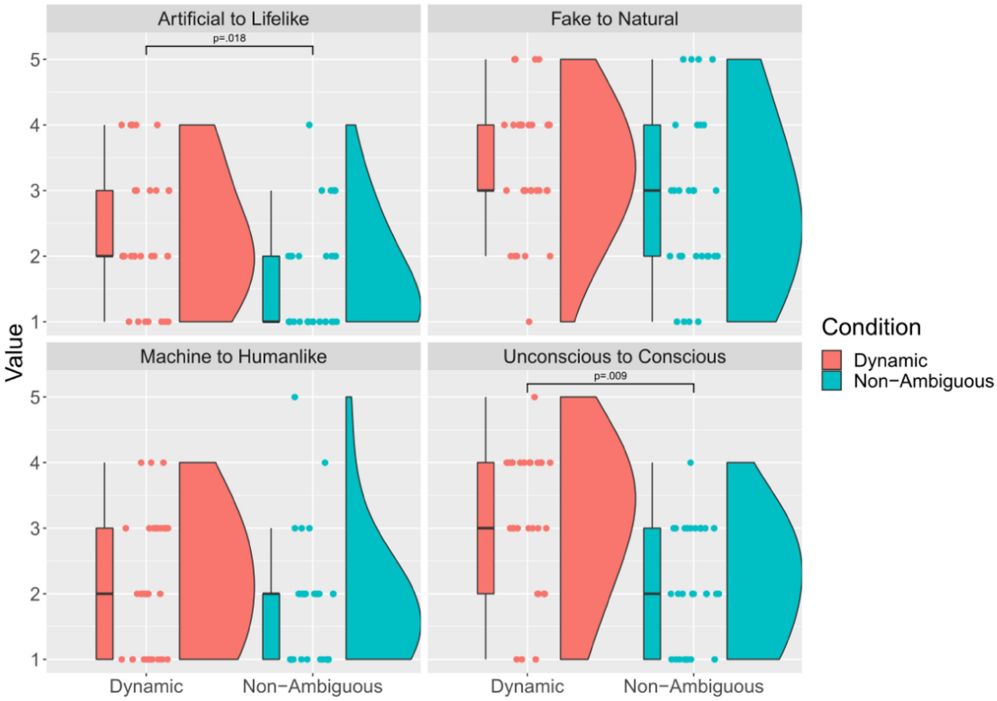

Fig. 10. Results of the questionnaire in the category of anthropomorphism.

participant felt that the cognitive load in the non-ambiguous was lower as they were having to "process potentially conflicting information".

3.3 Secondary Study

No significant difference was found between the two conditions on total time taken on questions (Welch two-sample t-test: $t = -0.3313$, $df = 46.985$, $p = 0.742$, mean of robot perspective descriptions = 425.70 seconds ($sd = 191.61$), mean of perspective-free descriptions = 445.24 seconds ($sd = 221.11$) (Figure 14). There was significance between the number of correct answers, with the perspective-free descriptions averaging higher scores (Welch two-sample t-test: $t = -2.212$, $df = 46.94$, $p = 0.032$, mean of robot perspective descriptions = 3.26 ($sd = 1.81$), mean of perspective-free descriptions = 4.46 ($sd = 1.98$) (Figure 15).

4 DISCUSSION

Overall, we see that the dynamic condition allowed participants to complete the task much quicker than during the non-ambiguous condition. The mean for the dynamic system was 62% that of the non-ambiguous condition. Unlike earlier studies, where we had seen this effect disappear as participants became more familiar with the task, this was maintained across both rounds of our task. This may be in part due to a much more complex task and with a higher cognitive load required to pilot the robot. However, we also need to address some of the issues that we had with the non-ambiguous descriptions, that may have made this task harder. The grey, silver and chrome barrels were confusing, and easily mixed up by participants, even with the tutorial at the beginning. Further, we found that the green barrels did not show up very well on the cameras, appearing black on the screens. This further added to the complexity of the non-ambiguous descriptions and may

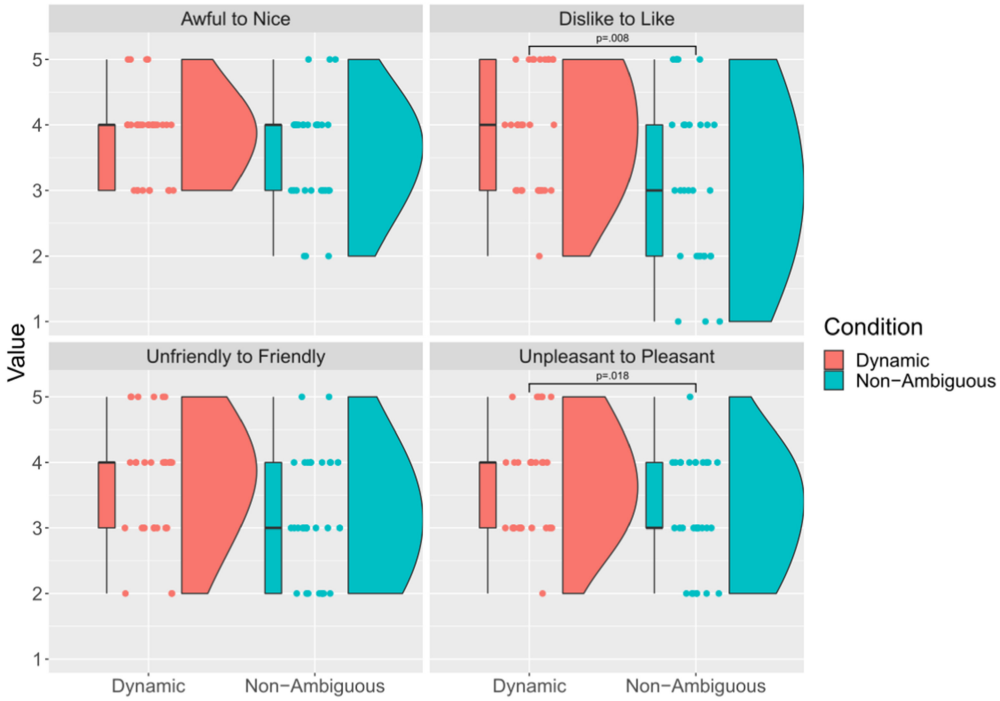

Fig. 11. Results of the questionnaire in the category of likeability.

have contributed to an otherwise non-ambiguous description being perceived as ambiguous. This only further shows how hard it is to create a non-ambiguous description, whereas the dynamic condition, while also suffering from the same issues, could help overcome this confusion. Still, the dynamic condition also requires accuracy of sensors, and for them to be reliable. Loss of readings from the Vive occurred for every participant at least once, and the contradicting information that arose from this issue led one participant to prefer the non-ambiguous communication.

We also see that the dynamic condition allowed participants to reach the target barrel much more directly, as we can see looking at the total distance travelled. We see a large distribution of the distance travelled when looking at the non-ambiguous condition. There are several reasons for this, the first of which is more often heading for the wrong barrel, whereas the dynamic condition would quickly correct participants. In addition, we saw that although they had a camera allowing them to see all the barrels, participants often preferred the robot's view to look directly at barrels.

We saw that for the most part, the dynamic condition was preferred. Participants mostly gave the reason of it being easier and more efficient. It was considered overall more helpful and the feedback more appropriate. Results from the questionnaire showed some improvements to the perception of how lifelike and conscious it seemed. The dynamic condition was also perceived as more knowledgeable, competent and intelligent. Although the timing of dynamic description was for the most part effective, there were still issues with the timing, especially on the amount of repetition by the robot. This may be what led one participant to feel the robot was too forceful. The timing here was based on a best judgement by us, but without participants being able to have a full two-way communication it was essentially guesswork as to whether the participants actually needed more information. Looking at being able to provide more natural timing on feedback remains an area where dynamic description could be significantly enhanced, along with participants

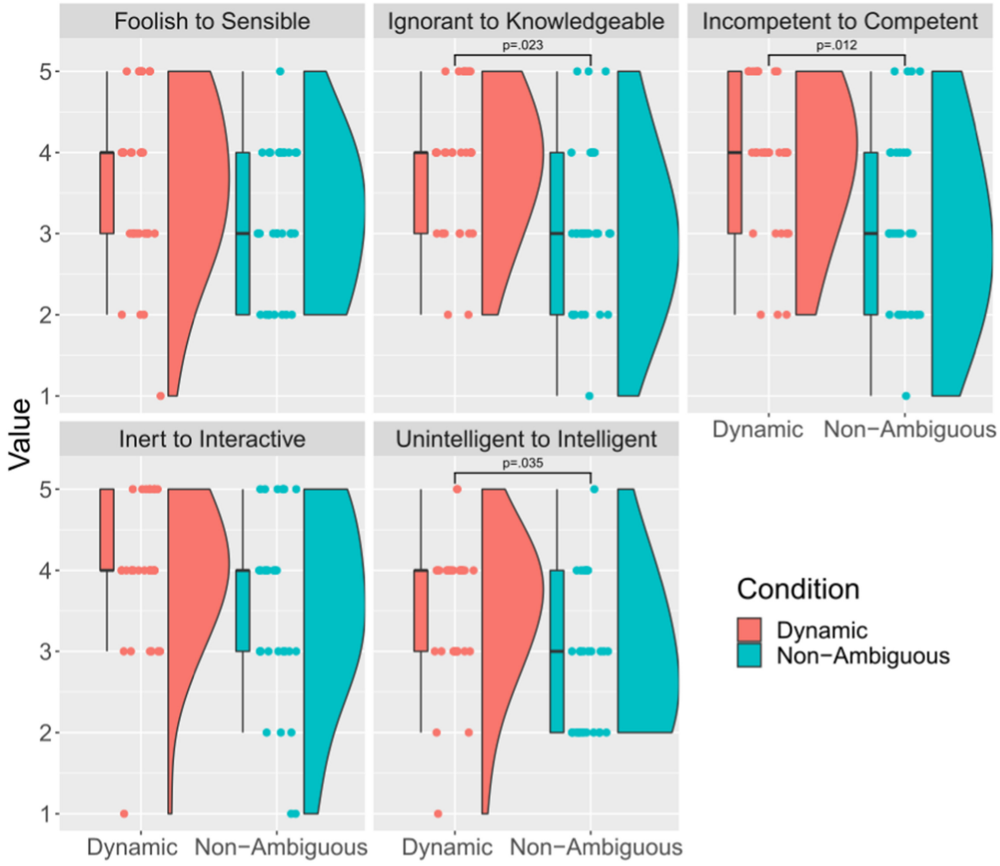

Fig. 12. Results of the questionnaire in the category of intelligence.

being able to provide spoken feedback to the system. Although we used an MLP based on previous work, a more thorough investigation of machine learning techniques may find other networks that may be more suited to producing natural timing.

Even though the locations of the barrels for the training were the same as used in the actual study, only positions relative to the robot were used to train the system. Localised non-ambiguous descriptions were generated based on observing the objects directly around the target—although we assume another system is providing labels for objects. Therefore, the robot would be able to generate descriptions based on many different configurations, both in a lab environment or in a real-world situation, without having to re-train the system. There are two limiting factors to this. The first is the robot is still only looking at objects on a 2D plane. Future work could look at describing objects that are at different heights. In addition, there was no need to navigate around other obstacles, as the robot just gave descriptions to targets in a straight line. More complex scenarios may need to look at how to describe objects that are not in immediate line of sight.

Although we have generated serviceable non-ambiguous descriptions, there is a question of whether these are the best descriptions we could generate, especially with the confusion with colours. This issue is one that has been seen previously in human-robot interaction, where robots and humans often have mismatched perceptions [17, 18]. Even in human-human interaction, however, what is said by one person is not always what is understood by another [14]. This

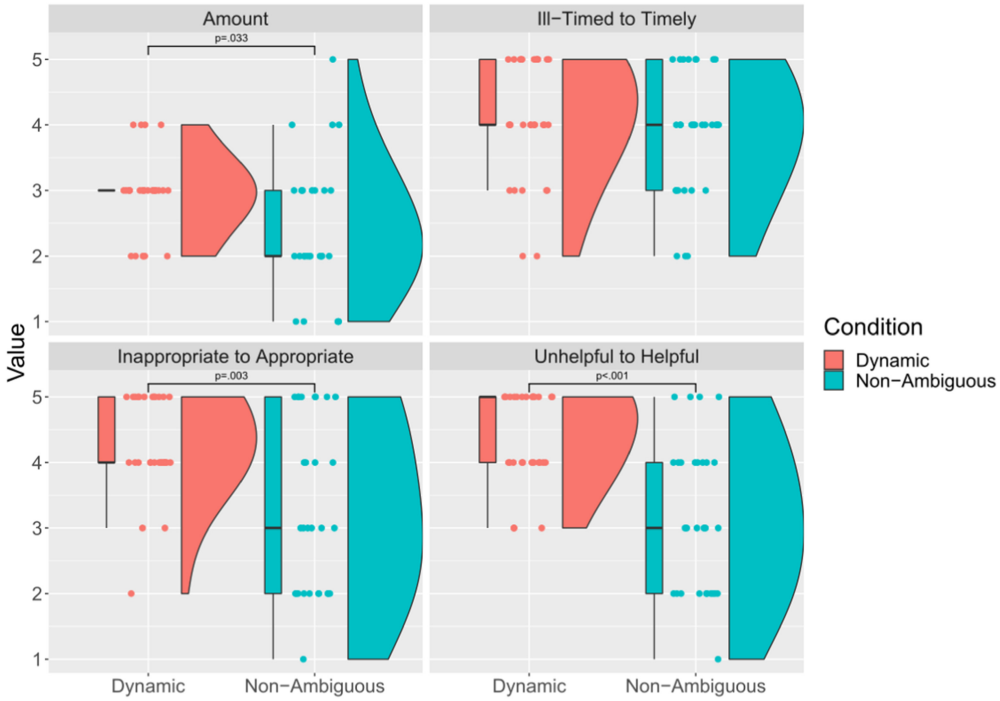

Fig. 13. Results of the questionnaire in the category of the feedback provided by the robot.

is usually resolved using dialogue to repair misunderstandings. Although not a true dialogue, with spoken two-way communication, the interaction provided by the dynamic system is able to resolve the issue in this case. However, in the situation where the robot was only able to repeat the non-ambiguous descriptions, participants were often confused as to what to do. Participants would often end up hesitating, or heading towards the wrong barrel or driving around to look at groups of barrels with the robot's camera, despite the overhead view.

Currently, we are using standard referring expression generation techniques to generate our non-ambiguous descriptions. However, it may be better to compare to the descriptions a human would make [31]. One potential avenue of research that may give us better non-ambiguous descriptions are machine learning based methods. Such algorithms are currently still focused on a fixed scene [8, 20]. In the real world, however, descriptions change as the situation does, and non-ambiguous descriptions during a live interaction are not a natural way for people to describe. To obtain human-generated non-ambiguous descriptions would require people describing a large number of pictures—to ensure a one way communication—that covered all of the situations that could occur. There would also be a question of whether such descriptions would stay consistent over time with such a system, or if they would fluctuate to completely different descriptions based on perspective. This may cause issues with finding objects. In which case the methods described in our dynamic system may be the solution -by defining what type of description should be given to keep a consistent description. We decided to run a study to find out if a perspective-based non-ambiguous description could in fact make a big difference to the time taken.

In our secondary study, we saw that there was no difference in time taken between using the perspective-free descriptions that we used in the main study and the description given from the robot's perspective, like the dynamic system ended up using. We do notice that the perspective-

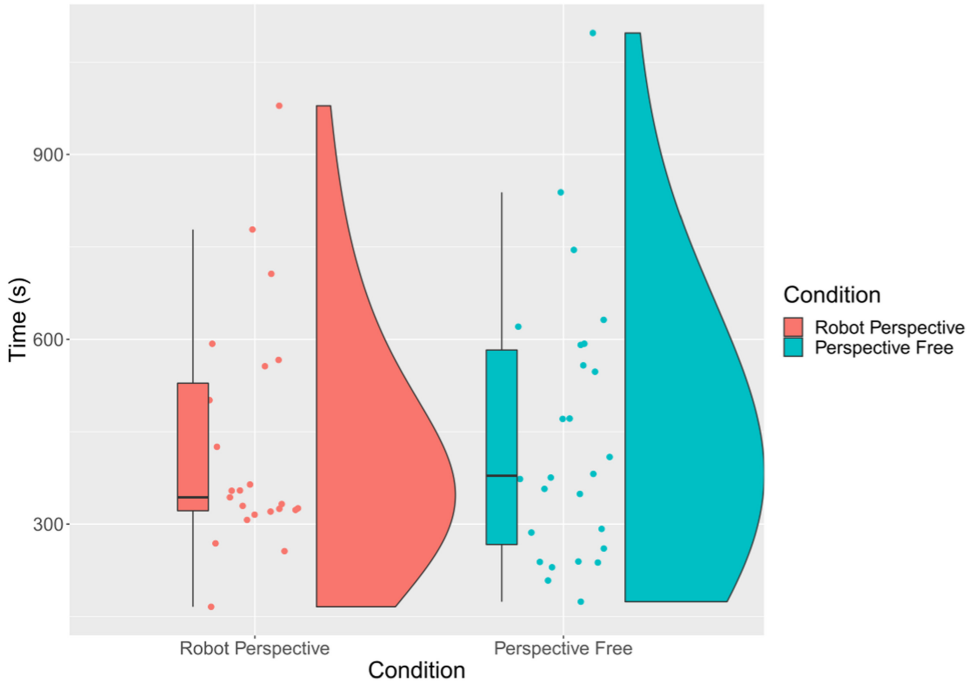

Fig. 14. Results of the time taken to answer all questions in the secondary study of robot perspective vs perspective-free. The difference in time taken was non-significant.

free descriptions result in a small increase in accuracy. In both cases, however, the accuracy was low, with the mean in both cases being under half of the barrels being correctly identified. This only strengthens the argument that generating good non-ambiguous descriptions in a complex environment is challenging, whereas incremental dynamic descriptions are easier to generate and likely to be more accurate.

There are also questions about the realism of this study. We focused on the process of actually generating the referring expressions, as opposed to how the robot actually acquired these targets. In this case, we assumed complete knowledge of the environment. This simplification would be considered the best-case scenario and was required to build non-ambiguous descriptions. A more realistic nuclear decommissioning scenario requires building maps of measured radioactivity. Therefore, at any given time, the state of the system is only partially observable, making it only more likely that dynamic descriptions would be required. As the information is gathered incrementally, so too would the robot need to dynamically update its instructions. Another simplifying assumption was that we would have a full and constant video stream from the robot. Although it would be desirable to somehow tag the object on the video stream, we would not be able to guarantee a constant live feed and may lose the feed on a temporary basis. The environment is highly radioactive, where thick concrete is used as shielding. There are likely to be areas where the bandwidth is lower. Instead, simple text instructions are much easier to transmit, and are likely to be received in a timely manner to support the piloting of the robot, than a full video stream, while still being able to view the robot itself from installed cameras in the environment.

Although the population group for the primary study was relatively close to our potential target group—skilled people trained to work with robots—there are likely still some differences. However, if anything, our assumption would be that robotics researchers are more likely to be biased towards existing techniques, which tend towards being non-ambiguous. Further studies should look at

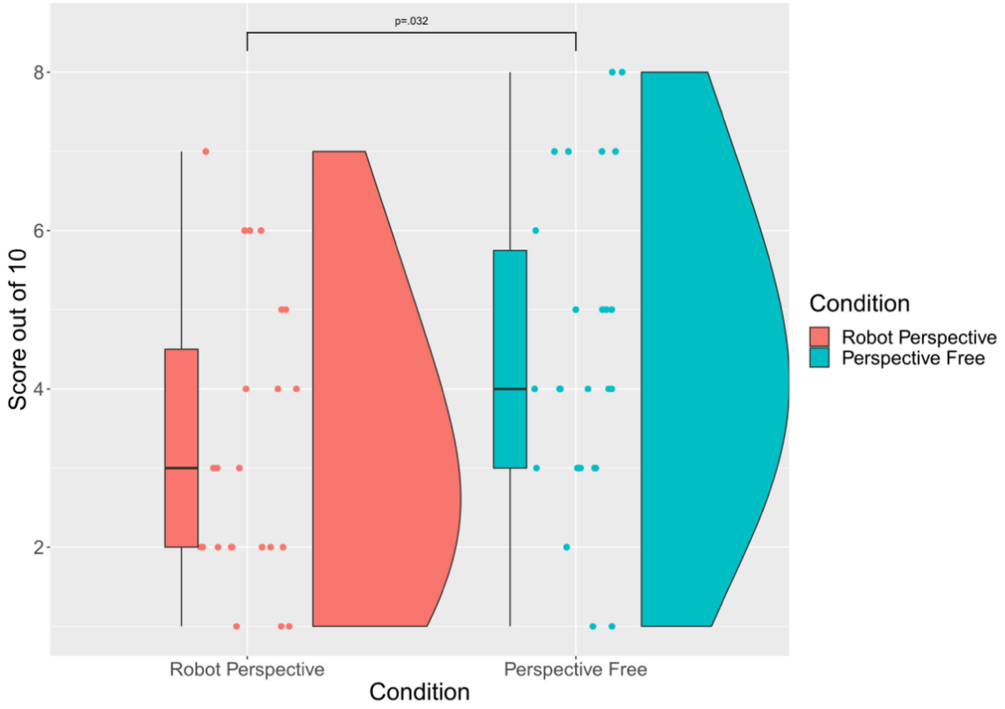

Fig. 15. Score of all questions in the secondary study of robot perspective vs perspective-free. The scores were significant, with those on the perspective-free condition being slightly more accurate.

more diverse population groups. These studies should also focus more on individual differences. Potentially by looking at people's skills in memory and spatial tasks, we may see more why some people may prefer a more non-ambiguous description compared to a more dynamic one.

5 CONCLUSION

We presented a dynamic system of generating spatial referring expressions for a potential real-world human-robot collaboration task, expanding from previous research on dynamic description. We tested the dynamic system in this task against standard non-ambiguous referring expression generation techniques. The dynamic system use an MLP network built by looking at two people performing the same task together. We used their utterances to classify the state of the interaction based on distance to target, change in distance to target, magnitude of motion from the previous position, yaw required, change in yaw and required change in yaw. As part of classifying these statements, we also provided a refinement to the topology of repair statements for interactive tasks first described in Wallbridge et al. [29].

In this remote piloting scenario, we found a significant reduction in the time taken to complete the task with our dynamic system of generation over that of the non-ambiguous. We also saw that in the dynamic system, the distance participants travelled to complete the task was much reduced. This supports our first hypothesis.

We also found that in a complex real-world scenario that a preference was shown for the dynamic description. With this system, and with the preference shown, we also see increased perception of its intelligence (in three categories: intelligence, knowledge and competence), likeability

(in two categories: likeability and pleasantness) and anthropomorphism (in two categories: lifelike and consciousness). This supports our second hypothesis.

In this scenario, we have a situation that is well suited to dynamic description. The robot is being piloted by the user, and simple contextual statements provide direct navigation instructions to control the robot. Future work on the dynamic system could look at a more collaborative scenario with more interaction between the robot and the user, and with them sharing the same space. We should also try to provide human-generated non-ambiguous descriptions as a better comparison point. However, these may be very challenging to generate, and it may be found that non-ambiguous descriptions are not sufficient for certain environments.

Improvements to the dynamic system could be made by looking at the timing of expressions in human-human interactions. More natural timing could lead to clearer instructions and less feelings of frustration generated due to the amount of repetition made by the robot.

REFERENCES

[1] GOV.UK. 2020. Direct Research Portfolio Annual Report 2018 to 2019. Retrieved August 23, 2021 from https://www.gov.uk/government/publications/direct-research-portfolio-annual-report-2018-to-2019/direct-research-annual-report-2018-to-2019.

[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81.

[3] Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science* 13, 2 (1989), 259–294.

[4] Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.

[5] Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19, 2 (1995), 233–263.

[6] David DeVault, Natalia Kariaeva Rutgers, Anubha Kothari, Iris Oved, and Matthew Stone. 2005. An information-state approach to collaborative reference. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. 1–4.

[7] Pierre Dillenbourg, Séverin Lemaignan, Mirweis Sangin, Nicolas Nova, and Gaëlle Molinari. 2016. The symmetry of partner modelling. *International Journal of Computer-Supported Collaborative Learning* 11, 2 (2016), 227–253.

[8] Fethiye Irmak Doğan, Sinan Kalkan, and Iolanda Leite. 2019. Learning to generate unambiguous spatial referring expressions for real-world environments. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'19)*. IEEE, Los Alamitos, CA, 4992–4999.

[9] Rui Fang, Malcolm Doering, and Joyce Y. Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. IEEE, Los Alamitos, CA, 271–278.

[10] H. Paul Grice. 1975. Logic and conversation. In *Speech Acts*, Peter Cole and Jerry Morgan (Eds.). University of California, Berkeley, 41–58.

[11] Peter A. Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics* 21, 3 (1995), 351–382.

[12] John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st COLING and the 44th Annual Meeting of the ACL*. 1041–1048.

[13] Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics* 38, 1 (2012), 173–218.

[14] Geert-Jan M. Kruijff, Miroslav Janíček, and Pierre Lison. 2010. Continual processing of situated dialogue in human-robot collaborative activities. In *Proceedings of the 2010 IEEE RO-MAN Conference*. IEEE, Los Alamitos, CA, 594–599.

[15] Séverin Lemaignan, Raquel Ros, E. Akin Sisbot, Rachid Alami, and Michael Beetz. 2012. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics* 4, 2 (2012), 181–199.

[16] Séverin Lemaignan, Yoan Sallami, Christopher Wallbridge, Aurélic Clodic, Tony Belpaeme, and Rachid Alami. 2018. Underworlds: Cascading situation assessment for robots. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)*. IEEE, Los Alamitos, CA, 7750–7757.

[17] Changsong Liu, Rui Fang, and Joyce Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 140–149.

[18] Changsong Liu, Rui Fang, Lanbo She, and Joyce Chai. 2013. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*. 78–86.

[19] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. 2018. A multimodal classifier generative adversarial network for carry and place tasks from ambiguous language instructions. arXiv:1806.03847.
[20] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. 2019. Multimodal attention branch network for perspective-free sentence generation. arXiv:1909.05664.
[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[22] Martin J. Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation* 4, 2-3 (2006), 203–228.
[23] Raquel Ros, Séverin Lemaignan, E. Akin Sisbot, Rachid Alami, Jasmin Steinwender, Katharina Hamann, and Felix Warneken. 2010. Which one? Grounding the referent based on efficient human-robot interaction. In *Proceedings of the 19th International Symposium on Robot and Human Interactive Communication*. IEEE, Los Alamitos, CA, 570–575.
[24] Mohit Shridhar and David Hsu. 2018. Interactive visual grounding of referring expressions for human-robot interaction. arXiv:1806.03831.
[25] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.
[26] James Temperton. 2017. Inside Sellafield: How the UK's most dangerous nuclear site is cleaning up its act. *Wired*. Retrieved August 23, 2021 from https://www.wired.co.uk/article/inside-sellafield-nuclear-waste-decommissioning.
[27] Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Natural Language Generation Conference*. 59–67.
[28] Christopher David Wallbridge, Séverin Lemaignan, Emmanuel Senft, and Tony Belpaeme. 2019. Generating spatial referring expressions in a social robot: Dynamic vs non-ambiguous. *Frontiers in Robotics and AI* 6 (2019), 67.
[29] Christopher David Wallbridge, Séverin Lemaignan, Emmanuel Senft, Charlotte Edmunds, and Tony Belpaeme. 2018. Spatial referring expressions in child-robot interaction: Let's be ambiguous! In *Proceedings of the 4th Workshop on Robots for Learning (R4L): Inclusive Learning @HRI 2018*.
[30] Tom Williams, Matthew Bussing, Sebastian Cabrol, Elizabeth Boyle, and Nhan Tran. 2019. Mixed reality deictic gesture for multi-modal robot communication. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI'19)*. IEEE, Los Alamitos, CA, 191–201.
[31] Tom Williams and Matthias Scheutz. 2017. Referring expression generation under uncertainty: Algorithm and evaluation framework. In *Proceedings of the 10th International Conference on Natural Language Generation*. 75–84.

Received January 2020; revised April 2021; accepted July 2021