2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 1-5, 2023. Detroit, USA

Development and Evaluation of Exploratory Experiences to Facilitate Reasoning About Robotic Systems

Sogol Balali, Marisa Hudspeth, Ian Afflerbach, Hannah Helgesen
Jessica McCurry, Walid Abu Al-Afia, Kelton Bruslind, Kathryn Hays
Ross T. Sowell, Ruth West, Cindy M. Grimm

2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-9190-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IROS55552.2023.10342409

***Abstract*— This paper introduces a novel interactive approach —Exploratory Experiences— that aims to improve the ability of people to reason about the capabilities and limitations of robotic technology. We focus on two areas: robot navigation and object detection. We evaluate the Exploratory Experiences with a novel approach that measures the participant's ability to predict when the robot will fail, following up with asking the *reason* and a possible *fix*. We show that our approach is effective at improving participants' understanding of potential robot navigation failures and that they already have the skills to detect potential object detection failures when presented with the correct stimuli.**

I. Introduction

Robots are appearing in public and semi-public places. Safely and effectively integrating these robots depends on a mix of factors, from robot engineering design to laws and policies that shape human-robot interactions, and how the public experiences and responds to them. It is unreasonable to expect all of the involved parties to have a deep technical understanding of how robots work. However, the lack of technical knowledge can lead to laws and policies that do not "make sense" in terms of what robots can (and cannot) do. In this paper, we take a first step towards a lightweight, interactive method for improving the ability of participants to reason properly about robotic capabilities and potential for failure. The goal is *not* to "teach" robot technology, but rather to let participants explore the technology "in action" by actively guiding them through where it succeeds and fails.

A fundamental challenge in understanding robotic capabilities is that people (even engineers) tend to anthropomorphize robots, in part because that is the closest mental model they have for characterizing robots [1]. Unfortunately, this leads people to believe that robots will sense, perceive, and take actions the same way that humans do. This tendency has been identified as "The Android Fallacy" [2], and it has deep implications for how robots (and their designers) are treated by the law. There are good reasons why robots/AI should *not* be treated as independent agents (e.g., humans) in most legal contexts [3]. The long-term goal of this work is to provide a lightweight approach to improving the ability of the law and policy community to reason properly about the capabilities (and limitations) of robots. By de-emphasizing the anthropomorphic aspects of the robot and focusing on specific failure modes, we can reduce the tendency to assume robots operate as people do.

In this paper, we focus on two areas of robotics (object detection and robot navigation — see Figure 1) that form the core of many robot applications. Within each area we provide just enough explanation of how the technology works for the participant to understand where (and how) it will fail — and guide them to produce those failures. The key to our approach is to provide interactive experiences that let the participant actually *cause* failures themselves. For example, the participant actually moves the camera to odd angles to interactively experience how object recognition often fails with those unusual angles. We call these combined explanations with interactive activities *Exploratory Experiences (EEs).*

To evaluate our Exploratory Experiences, we introduce a novel approach that measures a participant's ability to reason about a robot's capabilities and failures. Specifically, we provide stimuli (in the form of images — see Figure 2) and ask participants to choose which scenario(s) will cause a failure. For the predicted failures we follow up by asking the participants to pick the reason(s) for the failure, and then how to fix that failure (a form of data triangulation).

Our Exploratory Experiences are designed to be hands-on and interactive. In order to evaluate if interactivity is important (and also to provide web-based training materials), we replicate, as best as possible, the interactive sessions with videos, and organize the content into a website [1].

In summary, we develop novel reasoning-based interactive material and evaluate it with respect to the following hypotheses:

- H1: Exploratory Experiences can enhance people's ability to reason about the capabilities and limitations of robots.
- H2: Hands-on Exploratory Experiences are more effective than video-based ones in enhancing people's ability to reason about the capabilities and limitations of robots.

Funded in part by NSF grants NRI 2024872, 2024673, and 2024643.
Sogol Balali, Kelton Bruslind, and Cindy Grimm are with the School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, USA {balalis,bruslink,grimmc}@oregonstate.edu. Ian Afflerbach, Hannah Helgesen, Jessica McCurry, Kathryn Hays, and Ruth West are with the College of Engineering, College of Visual Art and Design, and College of Information, University of North Texas, Denton, USA {ianafflerbach,hannahhelgesen,jessicamccurry,kathryn.hays,ruth.west}@unt.edu. Ross Sowell is with the Department of Mathematics and Computer Science, The University of the South, Sewanee, USA rtsowell@sewanee.edu.

[1]Link to the Exploratory Experiences:
Object Detection: https://tinyurl.com/3u3xzs23
Robot Navigation: https://tinyurl.com/ms6hr2c4

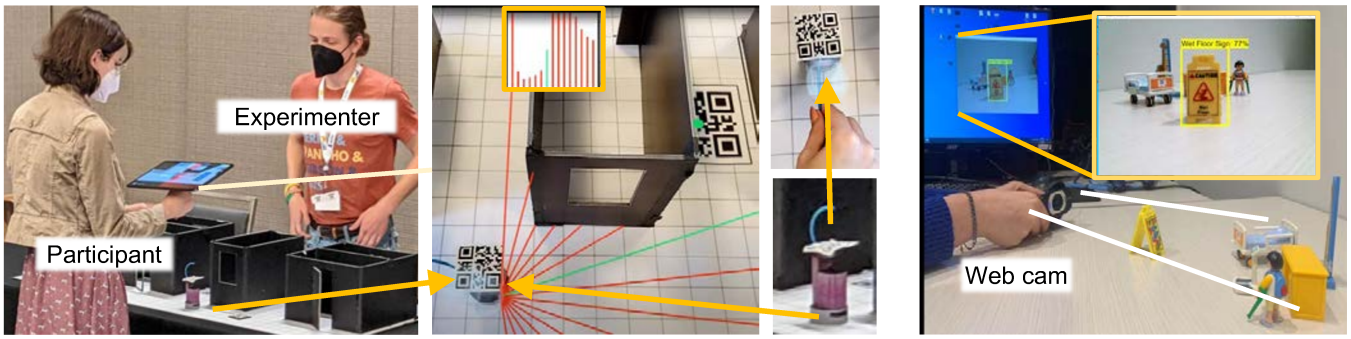

Fig. 1: Left: Hospital model (black) with a movable robot (pink with QR code at top) and IPad AR overlay (laser scan). As the robot is moved around the model, the IPad displays the laser plot (middle top) and (optionally) the laser beams. Right: Object detection — participant manipulates a web camera, objects, and lighting, to cause object detection (the yellow box around the wet floor sign) to fail.

Although not formalized as a hypothesis, we expect the learning gains to be greater for navigation (laser and external camera-based localization) than object recognition because people are broadly familiar with computer vision failures through, e.g., phone apps like Snapchat.

II. Related work

The development of EEs is situated at the intersection of multiple areas: design of explanations for eXplainable Artificial Intelligence (XAI) and explainable agency, interactive engagement in learning, evaluation of explanations, and post-training measures of change in understanding. We discuss these in turn.

A. Characteristics of effective explanations

Recent work in XAI [4] and explainable agency [5] offers criteria for explanations that effectively communicate the characteristics of AI and robotic systems. Kass et al. [6] propose basic principles for explanations of AI systems. We follow their construction in our explanations, with a particular focus on appropriateness, which refers to the adaptation of explanations based on the learners' knowledge. This principle is crucial in developing explanations for the EEs because the EEs are designed for, and need to be comprehensible to, people who lack a deep technical understanding of robot technology. We focus on describing causes for failure rather than oversimplified technical explanations because being too simple can result in lost trust in explanations [7].

Other works discuss various criteria for effective explanations. Kulesza et al. [8] contend that explanations should be sound, complete, and engage users' attention, while Mueller et al. [9] argue that explanations should correct users' oversimplified or reductive misunderstandings of AI systems. Further, works such as [10], [11], [12] emphasized the importance of contrastive explanations and maintained that explanations should explain the "Why", the "Why not", and the "What-if" of systems. Contrastive explanations are effective because they build understanding by highlighting the differences between events and are more easily comprehensible than a full causal analysis [13]. In this work, we used contrastive language to differentiate how humans perform tasks versus computer-based systems.

The work of [14] on producing effective explanations is the most closely related to this study. Their aim is to introduce explanations that enable people without technical understanding to identify the cause of, and solution to, a robot failure. They find that explanations should contain contextual reasoning about the environment and the history of a robot's past successful actions in order to improve failure and solution identification [14]. We follow a similar approach but define broader categories *a priori*, then guide participants through producing those failures.

B. Interactive engagement in learning

A limitation of lecture-based training is a lack of interaction and active participation in the process of learning, leading to a less effective conceptual understanding of a subject [15]. Wage et al. found that students in signals and systems courses learned only about 20% of concepts presented in a traditional lecture-based course format [16]. To address the limitations of traditional learning approaches in science majors, Hake [15, pp. 65] proposes the "Interactive Engagement" (IE) approach and defines it as methods "designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors." Hake found that IE results in greater conceptual understanding and problem-solving skills [15]. Our EEs incorporate IE through interactive activities that guide participants to generate failures and provide immediate feedback demonstrating failure occurrence.

C. Measures to evaluate explanations and understanding

Researchers in the field of XAI and explainable agency have developed several different measures for evaluating explanations [17][14]. Among such measures, failure and

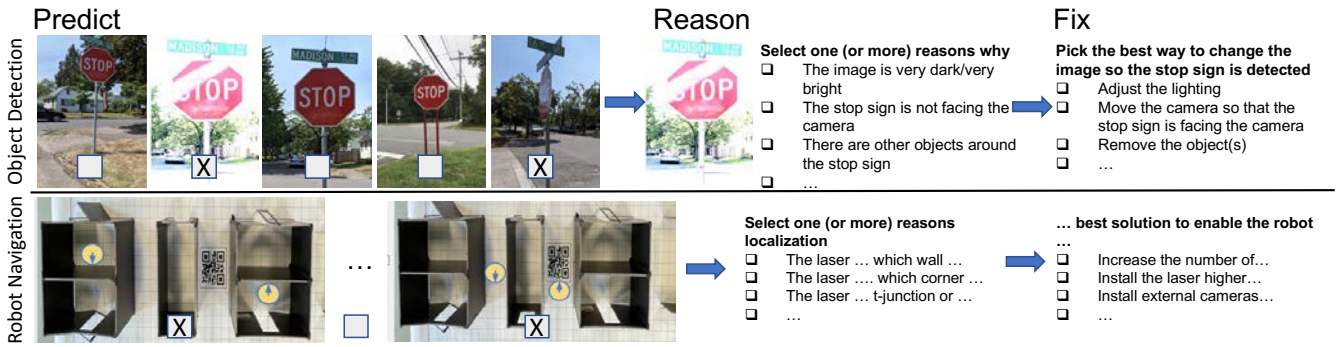

Fig. 2: Example Predict, Reason, Fix (PRF) questions for Object Detection (top) and Robot Navigation (bottom). Left: Participant selects one (or more) images that might fail. Middle: They select one (or more) reasons for the failure. Right: They select one best fix.

solution identification [14] are the most closely related ones to this work. This work introduces a new measure that considers not only a learner's ability to identify failures and fixes but also their ability to predict a failure.

To evaluate the effectiveness of different educational approaches, prior studies have collected learners' understanding of a subject before and after training through pre and post-tests [15], [18] or an assessment instrument known as concept inventory [19], [20]. Sands et al. [21] proposed a measure that evaluates the overall acquisition of new concepts after training, while Bristow et al. [22] introduce several measures that assess various aspects of a training approach, such as the areas where the training was ineffective and the extent to which it may have contributed to the development of misconceptions. In this work, we chose to use pre and post-tests to measure overall knowledge gain and targeted questions between each sub-task.

III. Methods

Each *Exploratory Experience* (EE) focuses on a specific robot technology and consists of a mix of short, non-anthropomorphic and non-technical explanations combined with a sequence of activities. The explanations cover the basic idea behind the technology and how it differs from the way humans do the same task. The activities involve manipulating the robot and/or its environment to produce specific outcomes (both successes and failures). Because driving a real robot around a real environment introduces too many unknowns and logistical challenges, we opt for a table-top scenario where the robot can be moved around a simplified environment (see Figure 1). We use Augmented Reality (AR) to show the participant what the robot's sensors are measuring.

To evaluate a participant's ability to reason about robotic capabilities, we present them with a set of scenarios and ask them to predict which ones might fail. The stimuli for these questions are a set of images that *show* the scenario (see Figure 2). To further evaluate the participant's ability to reason about robotic failures, we ask follow-up questions that ask them to pick the reason(s) for the failure and how the failure could be fixed.

We develop Exploratory Experiences for two areas of robotics: Object Detection (OD) and Robot Navigation (RN) (laser and external camera) (Section III-A). We use a website to organize the explanations and the guidance for each sequence of activities, along with the surveys. We also modify the websites to create a second, stand-alone website that replaces the interactive component with videos (Section III-B). Our study design is between-subjects, with an entry and exit survey and in-between surveys to capture participants' understanding after each activity (Section III-C). We use our novel Predict-Reason-Fix (PRF) measures to evaluate the effectiveness of the interactive versus web-based versions (Section III-E).

A. Exploratory Experiences: Content

Here we describe the technology each EE focuses on, along with the explanations and the physical setup the participant manipulates.

1) EE: Robot navigation: For robot navigation, we focus on two forms of localization (laser-based and external cameras) and object detection with lasers. Lasers are ubiquitous in robotics and are a good stand-in for any distance-based sensor (radar, lidar, optical flow). We include external camera triangulation for two reasons. 1) The technology is similar to GPS but at an interactive physical scale. 2) It introduces the concept that robot sensors do not have to be on the robot.

This study primarily focuses on robot localization (where is the robot on a given map?) and obstacle avoidance using the laser. In a preliminary study conducted around knowledge and understanding of navigation, it was clear that the public is very familiar with path planning through the use of map apps. For this reason, we excluded it from our sub-tasks for this evaluation. We leave map creation for future work. We next define the overall physical setup and the visualizations we created for each task.

Our navigation scenario is a simplified version of a hospital, with a couple of rooms with doors and windows (see Fig. 1, left), sized to fit on top of a table. The robot is a

EE	CONCEPT	SUB-TASK	TIME (m)	IN-BETWEEN	ENTRY/EXIT	TOTAL Qs	TIME (m)
RN	Navigation	Explanation	6	14 Q	2 Q	In-between: 28	43 ± 10
		Tutorial	6	3 P	-		
	Laser	Scan-location (I)	2.5	2 P	-		
		Uniqueness (I)	2.2	1 P	1 PRF		
		Limitations (I)	2.3	5 P	1 PRF	Entry/ Exit: 11 each	
		Obstacle (I)	1.6	1 P	-		
	External cameras	Tutorial	3	-	-		
		Localization (I)	1	2 P	1 PRF		
OD	Human vs computer Create fail (I)	Explanation	9	7 Q, 5 T/F	2 Q	IB: 27 E/E: 17 ea.	40 ± 7
		Distance, angle, light, Occluded, clutter	2-4	5 RFP	5 PRF		

TABLE I: EEs with sub-task type (I is interactive), time, and evaluation type: Q - Multiple choice, P - Predict, R - Reason, F - Fix, T/F - True/False. Entry and exit surveys had the same number of questions and took approximately 9 and 10 min each (half of experiment time).

3D-printed model that can be moved and tracked in AR within the hospital. We use augmented reality to simulate and visualize both the laser scan (the red lines in Fig. 3) and the field of view of "cameras" that are placed in the model (Fig. 4). For the AR view, we use a tablet that is held by the participant so they can see the hospital model; the image on the tablet is optionally shown on a monitor behind the hospital model. The tablet tracks what part of the model it is looking at in addition to where the robot is within the model.

In our preliminary study, we originally instructed the participant to move the robot while holding the AR tablet. This proved physically challenging; having the experimenter move the robot both resolved this problem and simplified the instructions.

Robot laser visualization: We have two modes for visualizing the laser. The first mode shows the lasers as red lines emanating from the robot in a 180-degree arc. The second plots the distances in a bar graph (see Fig. 3). After an initial explanation (once the participant indicates they understand the relationship between the robot's pose, the wall locations, and the laser scans), we turn off the first mode and only show the bar graph.

External camera visualization: We visualize the location and field of view of each camera using a blue dot plus lines (see Fig. 4). To emphasize that more cameras equals better localization, we draw a "ring of uncertainty" around the robot. The ring is bigger if fewer cameras see the robot, and goes from red (one camera) to green (three cameras).

We now describe the explanations and guidance we give for each of our three navigation EE.

Laser-based localization: After a brief explanation of what localization is, the participant watches a short video that explains how lasers work. The participant is then asked to watch how the bar graph changes as the experimenter: moves the robot forward and backward, rotates it in a hallway, and follows by a corner and a t-junction. The participant is asked to identify the differences between a bar graph at an intersection versus a hallway (Fig. 5, top row) to check that they understand before moving on. The participant next reads an explanation about how the robot can identify unique locations (the bar graphs are different). The experimenter then places the robot in two different locations that have the same bar graph (e.g., two hallways) to emphasize that those two locations look the "same" to the robot (e.g., corners, Fig. 5, bottom row).

Camera-based localization: The participant watches a short video that explains how one (or more) cameras can be used to locate an object using triangulation. The participant is then guided through placing the robot where it was well-localized (visible by all three cameras) versus not (one or zero cameras).

Object avoidance: After a brief explanation of how a robot could use the laser scanners to avoid running into an object, and how the laser could miss the object (e.g., wrong height, too skinny), the participant is guided through several cases where the lasers might "miss" the object. For these examples, the robot was placed in the model and moved toward the object in question (e.g., table with skinny legs, glass door) with the laser visualization turned on (see Figure 6).

2) EE: Object detection using a camera: In this EE, the participant interacts with a small "scene" consisting of a wet floor sign (the object to detect) along with potentially confounding objects, some of which are a similar color (see Fig. 1). We chose a wet floor sign because it has a 3D shape that changes based on viewpoint but is still "recognizable" from many viewpoints. The participant can move the camera, an optional light source (flashlight), and/or change the arrangement of the scene objects. A monitor shows the current camera image along with a labeled box if the sign is detected (no box means the wet floor sign is not detected).

The explanatory text first defines object detection (pixel values match pixel values from similar images, "detection" is the drawn box), then compares computer vision to human vision. The participant is next guided through five ways to make detection fail, with a short explanation of why that causes the failure (object too close/far, camera angle too far to the left/right/up/down, light too bright/dark, object occluded, clutter/confounding objects).

B. Websites and web-only versions

We use two Google Sites for organizing and presenting the Exploratory Experiences. The activities are ordered, with each activity consisting of the explanatory text followed by

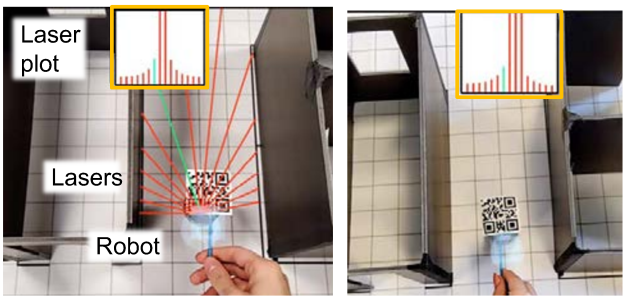

Fig. 3: Explaining the laser plot to participants. Left: We first show both the actual laser scans in the environment (red lines, one green to show left-right ordering) and the resulting distance plot (top middle). The laser scans move as the robot moves. Right: All subsequent interactions show only the laser plot.

the activity guidance, followed by one prediction and/or reasoning question. The first "activity" is the overall explanation of what that technology is. Bracketing the activities are the entry and exit surveys. Table I summarizes the activities in each.

To create our web-only versions, we replaced the guided activity instructions with a short video of a user performing that activity.

C. Overall study setup

Following pilot testing, we determined that an entire run-through of the Object Detection (OD) or Robot Navigation (RN) EE took around an hour (in-person or web-based). To prevent fatigue, each participant was randomly assigned to one of four conditions (in-person/web-based $\times$ OD/RN). The web-based versions were conducted remotely over Zoom to reduce the risk of COVID exposure. The in-person version was conducted in the lab, with the website on one monitor and the activities on a separate device. In all cases, following the consent process, participants began with an entry survey that collected familiarity with the given robot technology and asked a set of predict/reason questions, one for each activity in the EE. Following each activity, the participants answered an in-between survey before continuing to the next activity. The predict/reason questions from the entry survey were repeated in the exit survey, along with basic demographic questions and additional questions that asked the participants to evaluate the material and the effort it took to do the activities. A summary of all EE activities and questions is given in Table I. All questions are publicly available [2].

D. Participants

We recruited a total of 81 subjects (twenty for each condition), largely from university undergraduates and staff. Because of survey response failures, we had to exclude two, leaving 21 (OD, in-person), 19 (OD, web-only), 19 (RN, in-person), 20 (RN, web-only). Participants ranged in age from 18 to 50 years, with 74 participants in the 18-30 range, and

[2]Survey questions are here: `https://tinyurl.com/2p98jnes`

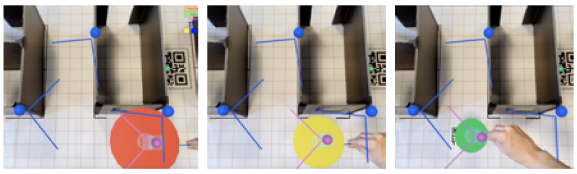

Fig. 4: Explaining localization using external cameras. From left to right: As the robot enters the camera's field of view (the blue circles with lines), a large red circle appears around the robot, illustrating the localization uncertainty. As more cameras see the robot (two then three), the circle shrinks and goes from yellow to green.

were approximately evenly split by gender across conditions (33 male, 42 female, 4 non-binary/declined). 38/40 (OD) and 32/39 (RN) were unfamiliar with the respective technology.

E. Measures and stimuli

Our primary method of measuring participants' ability to reason correctly is our *predict*, *reason*, and *fix* (PRF) score. Specifically, we show participants five images and ask them to select the ones that depict a potential failure (the Predict question). For the selected failure images, we ask a Reasoning follow-up question (select one or more reasons for the failure). This is followed by a Fix question (pick a single, best fix to address the problem). Both of the latter have an "other" option with a text box. For each of the five images, we calculate a score between 0 and 1, 0 being incorrect (specifics given below). The five images in each Predict question always consist of two "fail" images and three "not fail", with the images in random order.

In addition to the PRF score, we use traditional multiple choice questions (Q's) for conceptual questions such as "which tasks require navigation?". The question types are summarized in Table I.

In-between stimuli: We created one Predict question (no RF follow-up questions) for each sub-task in the Exploratory Experiences to serve as our evaluation for that sub-task's material (our in-between surveys). Because we were interested in reasoning, for Object Detection we also include one reasoning-plus-fix question for a (different) failure image. These images were made using the visuals in the EE (eg, with the wet floor sign in OD and the same viewpoint as the AR overlay in RN).

Entry-exit stimuli: For each sub-task, we created a different PRF to use in the entry/exit survey. For RN, these were also made using the AR overlay, but with an abstract picture of the robot (a yellow circle with an arrow indicating the direction it faced). For OD, we wanted to use real-world images (pictures of a stop sign) in order to see if participants could generalize. Although every effort was made to take two "fail" images for each failure type, and multiple, unique "not fail" images, several of the "not fail" images could be interpreted as "fail", and some of the "failure" images also had additional reasons they might fail.

Creating the code book/RF answers: We began by creating a list of common reasons why object detection and robot

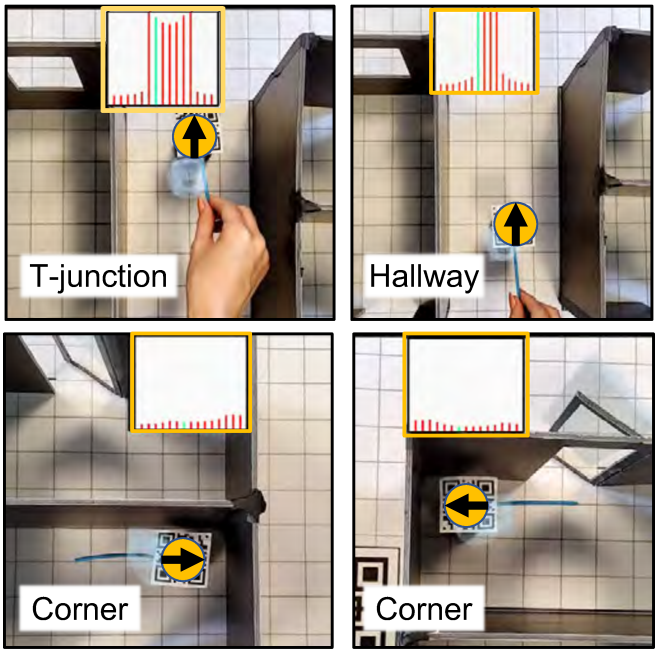

Fig. 5: Examples of location types and their corresponding laser plot. The robot's location and orientation are shown with a yellow circle and arrow. Bottom row: An example of demonstrating that pointing the robot into a corner produces the "same" laser plot.

navigation might fail. We use these to both select and design the EE sub-tasks and as the list of possible failure reasons. The fixes are actions that fix a specific failure case (e.g., fix the lighting, move the robot). After pilot testing, we collapsed a few of the reasons (e.g., clutter/graffiti, hallway versus corner localization) to reduce the overall study time. Three of the investigators reviewed the stimuli images to reach a consensus on the primary and secondary reasons to score failure images. This review confirmed that in Robot Navigation, only a few ambiguous images existed, and those show external camera localization (a robot seen by only 1 or 2 cameras).

Scoring a PRF question: Each Predict question has five images; we score each image individually and average the results. For each image, we calculate a prediction, reason, and fix score. The PRF score is the average of the three.

- Predict: 1 (correct) or 0 (incorrect).
- Reason: We use a modified multi-class F1 score, where only the primary reason must be selected; selecting correct secondary reasons is not penalized, but selecting incorrect ones are.
- Fix: 1 (correct) if the fix matched a selected failure and the failure was correct. 0.5 (semi-correct) if the fix matched a selected failure but the failure was incorrect. 0 (incorrect) otherwise.

The PRF scores are averaged to yield a score between 0 and 1 for each image. For each question, we average the PRF scores for each image.

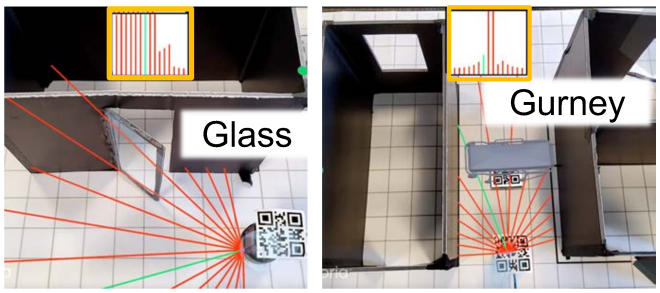

Fig. 6: Demonstrating that the laser scanner cannot detect transparent (glass door) and skinny (gurney legs) objects because the laser beams go through them.

IV. Results

We present results for each of our hypotheses (H1: effectiveness, H2: interactivity matters) for each type of Exploratory Experience, followed by demographics analysis and participant observations.

A. *Effectiveness*

We evaluate effectiveness both by comparing the entry and exit survey scores (Fig. 7) and by the in-between scores (Fig. 9). Overall, we saw significant improvement for Robot Navigation and high scores on the exit survey for both EEs.
Robot Navigation: There was a significant improvement from the entry to the exit survey for both the in-person and the web-only versions (PRF mean 0.75 to 0.9 for in-person, 0.78 to 0.82 for web-only). There were 28 questions total for the in-between surveys (14 multiple choice, 14 Predict, see Table I); the majority of the in-person participants missed 4 or fewer questions, while most of the web-only participants missed 5-6. Of the 14 Predict questions, the in-person participants missed an average of 1.2 questions while the web-only missed 2.7.
Object Detection: There were no significant differences between the entry and the exit surveys, largely because the participants scored high on the entry survey (0.86 mean both in-person and web-only). The exit survey scores declined slightly for the web-only condition (0.82 mean) but remained the same for the in-person.

There were 27 questions total for the in-between surveys (12 multiple choice or True/False, 5 RPF = 15 questions). Here, again, most in-person participants missed 5 or fewer questions, while most web-only participants missed 5 or more. Of the PRF questions, 15 people each in both conditions missed *none* of the RPF questions, with only one person in the online condition missing more than 2.

We hypothesize that the participants were already familiar with how well object detection works in practice through interaction with apps that use computer vision. Since the Reason and Predict scores were largely answered correctly in the in-between surveys, we performed an additional analysis on the entry/exit surveys to show the number of images participants selected as "failures" in the Predict questions (see Fig. 8). From this data we see that participants selected

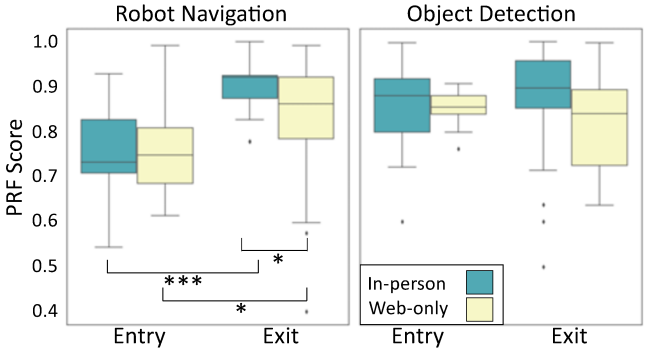

Fig. 7: PRF scores before and after the EE (Robot Navigation, 3 PRF questions total, Object Detection, 5 total). ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$).

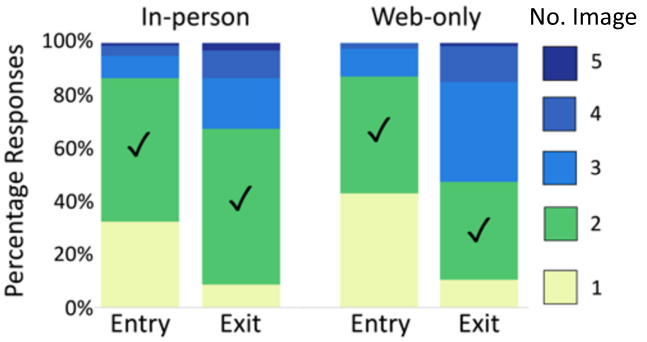

Fig. 8: Number of images predicted to be failures in the Object Detection prediction questions. Two of the 5 images were "fails". For both in-person and online, more participants selected more images per question *after* the EE (Total: 5 prediction questions per participant).

more failure images in the exit survey; this was particularly pronounced for the web-only participants, who shifted from selecting 1 failure image to 3 (the "correct" number was 2). We hypothesize that the EE resulted in participants being hypersensitive to potential failures. Note that we used real-world images for the entry and exit surveys, so some of them could be interpreted as, e.g., "too dark". Although we do not include the analysis here because of space, we did re-code the "successful" images using a lower bar for failure and discovered that participants were, indeed, marking these potential fails with the reasons we assigned to the images.

B. Interactivity

The web-only participants performed worse both in the exit surveys (Fig. 7) and the in-between surveys (Fig. 9). For Object Detection, the web-only participants did *worse* on the exit survey (although this was not statistically significant).

C. Demographics and follow-up questions

In the exit survey, we collected demographic information (age, gender, attitude towards technology), a 6 question workload survey, and 6 questions about course content. While our participant pool was roughly evenly split by

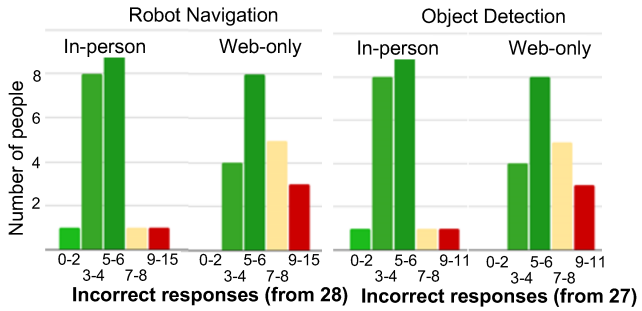

Fig. 9: Number of incorrect answers for the in-between surveys by EE.

gender, our participant pool skewed young (nearly all in the 18-30 range). Because of this, we did not run age-based correlations.

We saw no statistically significant difference between the male and female participants on any of the entry surveys in any of the conditions. We did see a **gender** difference in the Robot Navigation web-only condition (males performed better), both in the exit survey (males: mean PRF 0.90, females: mean PRF 0.75, p-value 0.031) and in the in-between surveys (males: missed 4.7 questions on average, females 8.9, p-value 0.008). This was not, however, the case for the in-person Robot Navigation, where there was no statistically significant difference between the scores in the exit or in-between surveys. We also saw a gender difference in the Object Detection in-person condition, but in the other direction (females performed better) in the in-between surveys (males: missed 6.12 questions on average, females 4.2, p-value 0.041). Altogether, this indicates that the in-person training was more effective for females than males.

We asked the NASA TLX questions (paper version, 7 questions, 21-point Likert scale, 1-low, 21-high) at the end of the experiment to evaluate perceived effort and performance. Scores were very similar across all conditions. The exception was temporal demand (feeling rushed), which was noticeably higher for the web-only Robot Navigation than the in-person (avg. 3.3 vs. 0.95, p-value 0.03). Overall, participants felt they were successful (avg. 15.6) with higher mental demand (avg. 4.7) than physical (avg. 1.1), with low effort (avg. 4.4) and frustration (avg. 1.7). This correlates with their responses to the course content, where the majority of the participants found the content useful (avg. 4.5/5, std. 0.90), sufficient (avg. 4.5/5, std. 0.92), and easy to understand (avg. 4.4/5, std 0.95). The testing questions were also easy to understand (avg. 0.45/5, std. 0.88). The Robot Navigation, versus the Object Detection, scores were slightly higher (0.3 on average), indicating that the Robot Navigation experiences were slightly more engaging.

We found no correlations between performance and our two technology questions (comfort with technology, whether or not they like to tinker). Our participants did skew towards being comfortable with technology and willing to tinker with it. When asked if the content in the course was new, 26 of the

Object Detection participants agreed (as compared to 38 who said they were unfamiliar with object detection technology in the entry survey). For Robot Navigation, this was 27 and 32, respectively.

V. Discussion

Our goal was to determine if interactively engaging with robotic technology — in particular, deliberately causing it to fail — is an effective method for improving the ability of participants to reason about capabilities and potential failures. Our results for Robot Navigation showed that this form of failure-based explanation was effective. More intriguing is how well participants did on the *entry* surveys — even the Robot Navigation, which does not have an equivalent app (like face recognition) that participants are exposed to in daily life. It is clear that participants have a pretty intuitive understanding of how object recognition might fail *if* you show them stimuli (images) that exhibit common failures. Similarly, we think that participants — knowing nothing about how laser scans or external camera localization works — can similarly reason from physical scenarios, such as a robot pointed at two similar corners. This suggests that grounding robot capabilities in physical scenarios may work better than generic questions such as "when do you think a robot might get lost?"

It was clear from the evaluations (and participants' comments) that in-person, interactive was more effective, even for exploratory content that was largely delivered via text on the webpages (the non-PRF in-between questions). This might, in part, be attributed to the physical presence of the experimenter (as opposed to a virtual Zoom presence), which caused them to pay more attention to the content. However, the effect for Robot Navigation was sufficiently strong to suggest that physically causing the failures (rather than passively observing them) is more effective. The decline in scores for the web-only Object Detection condition (but not the in-person) could also have been caused by boredom and general fatigue.

Based on the success of the hands-on exploratory experiences in the controlled experiment presented in this paper, we brought the demos to two conferences (We Robot 2022 in Seattle and Science Writers 2022 in Memphis) to try them out in a more realistic, informal setting. In both cases, we were set up in an exhibit hall where conference attendees could stop by for a few minutes during breaks and try one or both of the EEs. The different conferences gave us access to two different audiences. We Robot attendees tend to be lawyers and policymakers in the technology space. The Science Writers attendees are journalists who specialize in writing about science. While both groups are highly educated professionals, neither has particular technical training in robotics. While the informal nature of this setting made it difficult to collect robust data, the feedback that we received was overwhelmingly positive. People enjoyed the hands-on exploratory experiences, and we have now had over one hundred people interact with them. We did ask people to voluntarily complete a brief two-question survey on a clipboard after the EE, and the results from this indicate that people were able to successfully identify failure cases after just a brief interaction with the EE.

Improving reasoning ability is the first step toward our long-term goal. In future work, we plan to study if interacting with these EEs changes how people evaluate potential laws and policies, particularly around technology such as sidewalk robots.

References

[1] M. Goodrich, "Using models of cognition in hri evaluation and design," 10 2011.

[2] N. M. Richards and W. D. Smart, *How should the law think about robots?* Cheltenham, UK: Edward Elgar Publishing, 2016.

[3] R. Calo, E. Kumar, A. Selbst, and S. Venkatashubramanian, "The legal construction of black boxes," in *We Robot 2021*, 9 2021.

[4] D. Gunning, "Explainable artificial intelligence (xai)," 2018.

[5] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," ser. AAAI'17, 2017, p. 4762–4763.

[6] R. Kass and T. Finin, "The need for user models in generating expert system explanations," vol. 1, 10 1988.

[7] 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 *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, Sep. 2013, pp. 3–10.

[8] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," vol. 2015, 03 2015.

[9] S. T. Mueller, R. R. Hoffman, W. J. Clancey, A. Emrey, and G. Klein, "Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI," *CoRR*, vol. abs/1902.01876, 2019.

[10] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2280–2288.

[11] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016.

[12] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[13] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplements*, vol. 27, p. 247–266, 1990.

[14] D. Das, S. Banerjee, and S. Chernova, "Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery." New York, NY, USA: Association for Computing Machinery, 2021.

[15] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American journal of Physics*, vol. 66, no. 1, pp. 64–74, 1998.

[16] K. Wage, J. Buck, C. Wright, and T. Welch, "The signals and systems concept inventory," *Education, IEEE Transactions on*, vol. 48, pp. 448 – 461, 09 2005.

[17] S. Mohseni, N. Zarei, and E. D. Ragan, "A survey of evaluation methods and measures for interpretable machine learning," *CoRR*, vol. abs/1811.11839, 2018.

[18] I. A. Halloun and D. Hestenes, "The initial knowledge state of college physics students," *American journal of Physics*, vol. 53, no. 11, pp. 1043–1055, 1985.

[19] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," *The physics teacher*, vol. 30, no. 3, pp. 141–158, 1992.

[20] R. Gerndt and J. Lüssem, "Towards a robotics concept inventory," in *6th International Conference on Robotics in Education. Yverdon-les-Bains, Switzerland*, 2015.

[21] D. Sands, M. Parker, H. Hedgeland, S. Jordan, and R. Galloway, "Using concept inventories to measure understanding," *Higher Education Pedagogies*, vol. 3, no. 1, pp. 173–182, 2018.

[22] M. Bristow, K. Erkorkmaz, J. P. Huissoon, S. Jeon, W. S. Owen, S. L. Waslander, and G. D. Stubley, "A control systems concept inventory test design and assessment," *IEEE Transactions on Education*, vol. 55, no. 2, pp. 203–212, 2012.