2023 32nd IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 28-31, 2023. Paradise Hotel, Busan, Korea

2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) | 979-8-3503-3670-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/RO-MAN57019.2023.10309524

Towards Improving User Expectations of Robots by Leveraging Their Experience With Computer Vision Apps

Sogol Balali, Ian Afflerbach, Ross T. Sowell, Ruth West, Cindy M. Grimm

***Abstract*— This paper explores whether experiential knowledge of computer vision from interacting with daily apps (e.g., Instagram, Zoom, etc.) can be leveraged to improve users' expectations of robotic capabilities. We evaluate users' ability to predict when computer vision apps might fail and if they can apply their experience to reason about computer vision in robotic systems. We show that although users can reliably predict computer vision app capabilities and functionality, they tend to ascribe human-level knowledge to those apps and do not reliably correlate app functionality with similar robotic tasks. We propose that experiential knowledge gained through interaction with software apps is a potential way to "calibrate" user expectations of the function and failure states of complex systems.**

I. Introduction

Intelligent robotic systems employ fundamentally different sensors, actuators, and algorithms than people do, yet the distinction between human intelligence and intelligent systems is often blurred by anthropomorphization. This mismatch causes problems because we (researchers, developers, application designers, and policy makers) blithely assume that the two systems will succeed (and fail) in the same way and for the same reasons. As an example, robots do not "see humans". They have laser and camera sensors that result in a set of measurements that are then compared to previously seen data that has been labeled as "person", and combined with a set of (often implicit) assumptions. If the match is "close enough" (in mathematical terms) a box is drawn around the area in the image and labeled as "person". This mathematical calculation can fail, leading to not detecting a person (or detecting one where there isn't) — in situations no human would fail at the same task.

In this study, we explore whether interaction with daily software applications (e.g., Instagram, Zoom) that use Computer Vision (CV), can enable users to reason properly about the functionality, capabilities, and limitations of such systems. Moreover, we investigate if this knowledge can be used to identify the same technology when embodied in a robotic system.

To explore participants' understanding of everyday software applications, we ask them questions about the functionality, capabilities, and limitations of several widely-used CV software applications. We assess whether the nuanced differences between the apps are reflected in participants' responses. We also measure how well they can predict when an app might fail, and how well they can identify the cause(s) of, and solution to, the failure. We ask them questions about functionality to determine if they understand the difference between how a human might do the task versus what the algorithm does. To learn whether their experiential knowledge of everyday software applications helps them recognize the same technology in robots, we ask them to rate the extent to which they believe a task-specific robot (e.g., receptionist robot) uses a technology similar to the ones used in the everyday app.

This topic is important because prior literature maintains that the understanding that people develop through interaction with robots about their capabilities and limitations may be inaccurate, reductive, or involve potentially harmful simplifications [1]. Such understanding can have serious consequences, including loss of trust [2] in a robotic system or even disuse of it [3]. Our work aims to understand how we might leverage people's experience with everyday apps to improve their ability to reason about robotic systems.

In this research, we focus on computer vision because it is widely used in robots that need to operate in the real world, and people have some experience with it through apps that they already use in their daily lives. Our research questions are: (RQ.1) What features within the image do participants believe that these apps use to do their tasks?; (RQ.2) How well can participants reason about the capabilities and limitations of these apps?; (RQ.3) Can participants identify the same technology when it is embodied in a robotic system?

Our study measures participants' experiential understanding (can they predict when the app will succeed/fail, and what might cause those failures) as well as what types of features (eg, noses, hair color) these apps might use. We show that participants can effectively do this prediction, but they tend to ascribe human-type knowledge (the app "knows" there is a face) to the algorithm. Interestingly, they have not made the connection that this type of technology is what robots might use to perform their tasks.

II. Related Work

We discuss work in interactive machine learning and explainable artificial intelligence.

A. Interactive Machine Learning

The term Interactive Machine Learning (IML) was first introduced in 2003 [4] to address the key limitations of

Funded in part by NSF grants NRI 2024872, 2024673, and 2024643. Sogol Balali and Cindy Grimm are with the School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, USA {balalis,grimmc}@oregonstate.edu. Ian Afflerbach 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,ruth.west}@unt.edu. Ross T. Sowell is with the Department of Mathematics and Computer Science, The University of the South, Sewanee, USA {rsowell}@sewanee.edu.

APP TYPE	DESCRIPTION	EXAMPLE
Filter *	identifies a face and places virtual face or background filters on or around the face.	Instagram, Snapchat, Zoom, Skype
Tag	recognizes and tags people in images.	Facebook, Google Photos
Lock *	recognizes a face to unlock the device.	Smartphones, smart door locks
Text *	scans a word to get an instant translation in a different language.	Google Translate
Check *	scans a check to deposit it to a bank account.	Online banking apps
Food	scans a plate of food to return the corresponding calorie of the items on the plate.	Calorie counting apps
Finger-lock	scans a fingerprint to unlock a device.	Smartphones, smart door locks
Barcode	scans a barcode of a product to find the product information on a retailer website or find nutrition information.	Amazon Barcode Scanner, MyFitnessPal

TABLE I: List of all apps initially considered for this study. The four marked with a * are the ones selected for this paper.

Classical Machine Learning (CML) Models. These limitations include slow training time and the absence of corrective feedback from users during the training process. The IML approach [4], [5] incorporates quick train-feedback-correct cycles that enable users with no machine learning background to rapidly correct the mistakes made by Machine Learning systems. The IML approach also allows users to adapt their own feedback behavior based on the system behavior and even learn from it. IML research is relevant to our study in that gaining understanding of Machine Learning systems through interaction with them (i.e., experiential knowledge), especially for end users, is a key component in that area of research. In our work, we focus on everyday apps that use computer vision. Unlike the IML literature, we investigate the impact of experiential knowledge on users' ability to reason about the system functionality, capabilities, and limitations, not to improve the system's performance.

The study by Kulesza et al. [6] is perhaps the most relevant to our work because they investigate the impact of interaction with IML systems on people's ability to build useful mental models [7] of such systems. In their research, their main measure is the users' ability to personalize the system. In our research, we instead measure the users' ability to predict system behavior.

B. Explainable Artificial Intelligence (XAI)

XAI can be defined as a self-explanatory intelligent system that describes the reasoning behind its decisions and predictions [8]. The concept of XAI can be applied to any specific sub-field of AI, including computer vision. For instance, eXplainable Face Recognition (XFR) focuses on explaining why a face-matching system matches faces. Studies in this area introduced comprehensive benchmark evaluation for XFR, providing ground truth in order to quantify the image regions that contribute to face matching [9], and proposed approaches such as using visual psychophysics to make face recognition algorithms more explainable [10].

XAI literature studies several key topics concerning XAI, such as metrics to evaluate XAI, and design guidelines for XAI [8]. However, to date, users' understanding obtained through their prior experience (i.e., experiential knowledge) with AI applications has not been considered a key factor in the design of XAI.

XAI also suggests various criteria for explanations to effectively communicate the characteristics of AI systems. For instance, several researchers [11], [12], [13] believe that explanations should be contrastive, that is they should explain the "Why", the "Why not", and the "What-if" of systems. Contrastive explanations are more effective than full causal analysis because they focus on emphasizing the differences between events [12].

Explanations also should be sound, engaging, and corrective to address users' reductive or oversimplified understandings of AI systems [6]. Explanations must enrich mental models but also correct user misunderstandings [1]. The findings of this study provide a base-level understanding of users' mental models that can be used to develop corrective and contrastive explanations in XAI more effectively.

III. Methods

We designed an online survey [1] to evaluate the effectiveness of experiential knowledge in enabling people to reason about the functionality, capabilities, and limitations of everyday software applications that use computer vision technology. We distributed the survey using Amazon Mechanical Turk in order to recruit participants from a relatively large population with different levels of familiarity with these applications.

We started with 8 widely-used software applications (see Table I). After pilot testing, we narrowed this list down to 4 apps that differ in two dimensions: face (i.e., apps using facial features) versus text (i.e., apps using features of texts), and class (e.g., detecting any face) versus instance (e.g., detecting a specific face). We refer to these 4 apps as *Filter*, *Lock*, *Text*, and *Check* (marked with * in Table I).

More specifically, we used the following criteria to pick the 4 applications: 1) *Familiarity*: apps that the general public is likely to be highly familiar with and/or use regularly. 2) *Interactivity*: apps that allow users to interact actively and make live changes (e.g., change the lighting or the camera's orientation) to improve the app's performance.

Each of the **Face-based** and **Text-based** categories consists of one application that does detection (e.g., detecting any face) and one application that does recognition (e.g., detecting a specific face). The goal is to see if participants understand the nuanced differences between apps that appear to have similar (e.g., processing facial features) but, in essence, different functionality (e.g., one app detects any face while the other app recognizes a specific person). Therefore, the **Face-based** category consists of *Filter* (detection) and

[1]Survey questions are here: `https://tinyurl.com/mpeetve9`.

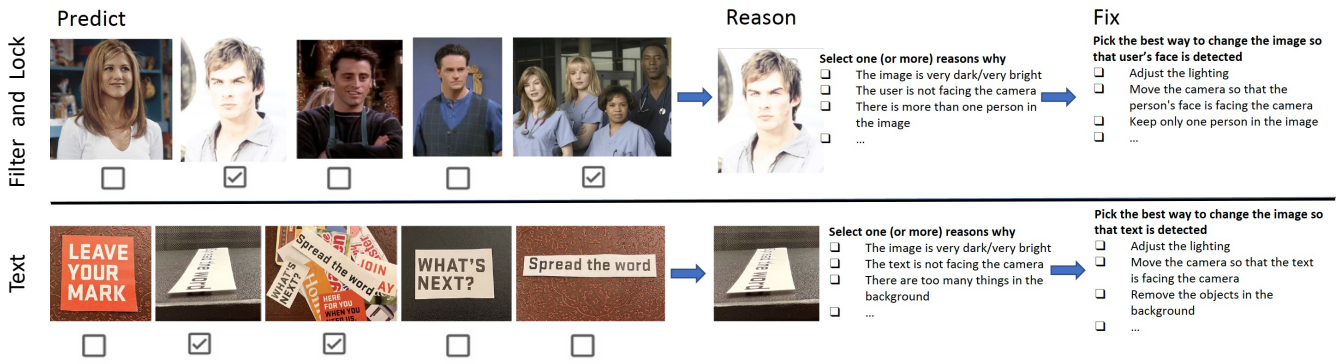

Fig. 1: Example *Predict*, *Reason*, *Fix* questions for *Filter* and *Lock* (top) and *Text* (bottom). Left: Participant selects one (or more) images that might fail. Middle: They select one (or more) reasons for the fail. Right: They select one best fix.

Lock (recognition), and **Text-based** category consists of *Text* (detection) and *Check* (recognition).

A. Survey design

The survey has three sections: App familiarity (used to ensure that the participants had actually used the app), Application function/capability questions (9 questions for each application), and a short demographic section. The survey took, on average, 19.11 minutes. We included one attention check question per application.

Each participant was randomly assigned to two apps of the four (either the two **Face-based** or the two **Text-based**, order randomized) to prevent fatigue. A participant was only assigned to the apps they were familiar with. In a few cases a participant was only familiar with one app in one of the two categories; they only answered questions for that app.

9 application questions: Because we wanted to compare responses across apps, these 9 questions were chosen to be as similar as possible while reflecting the different apps (e.g., the specific features we use differ — see Table II). The specific questions we ask are: Q1) *Knowledge source* where participants learned how the app functions (interaction, reading documentation, other people), Q2) *Features* what features the app uses to do its task (Table II), Q3-Q7) *Capabilities* what capabilities the app has, expressed as a level of agreement with statements (Table III), Q8) *Predict* is a multi-part question that asks the participant to predict which images might produce failures and why (Fig. 1), Q9) *Similar technology* asks the participant which of several robotic tasks might use similar computer vision technology.

Our *Predict* question (Q8) is a multi-part question designed to capture prediction ability using stimuli that matches what a participant would see in a typical use case. Fig. 1 shows the question flow and the types of images. We show 5 images and ask participants to pick the one(s) that depict a potential failure. For each selected image, we follow up with a *Reason* and a *Fix* question. The *Reason* question asks the participant to pick the reason for the failure — e.g., too dark, and the *Fix* to choose a way to fix the failure — e.g., turn a light on. The fixes were matched to the failures.

For the *Predict* question designed for each app, we created multiple sets of 5 images and randomly assigned the participant to 2 of those sets. Each set had 2 images that would "fail" based on common criteria (too dark/light, bad camera angle, too close/far, occluded, background clutter), and 3 that were "not fails". The face images were selected from well-known television series (e.g., Friends, Lost) and were selected to have a diverse set of skin tones and face shapes, but still be real-world images that contain human faces. The text/check images were created by taking pictures ourselves because it was easy to create diverse images by changing different features of the image (e.g., changing backgrounds, adding different objects).

For the *Capability* questions, we started with a set of statements that expose typical anthropomorphic-based misunderstandings of how current state-of-the-art computer vision algorithms function. Two investigators reviewed these statements until they reached an agreement on the clarity of each question and the "correct" answer (marked as Agree or Disagree in Table III).

For the *Similar technology* question, we started with a list of potential near-term robot applications/tasks, identifying four that would require similar technology (see Table V).

We used a 5-point Likert scale for Q1, 3-7, and 9, which were collapsed in the plots to a 3-point to improve readability. Q2 (Feature) used a 3-point Likert scale. Q8 (Predict) used multiple-choice for the Reason and Fix questions; both had an optional "other" option with a text-box answer.

B. Participants

We recruited 68 participants in total. We excluded the responses (*Filter* and *Lock*: 3, *Text*: 9,*Check*: 7 responses) of participants who failed the attention check question. Table IV shows the number of participants per application type and their demographic information.

C. Data Analysis

For the *Capability* questions, we show the percentage of participants agreeing, disagreeing, or neither for each question and each app; the correct answers are marked with check marks on Fig. 6.

APP	FEATURE	EXAMPLE
Face-based	Unchangeable	nose, eyes, lips, face shape
	Changeable	eyebrow shape, hairstyle, facial hair
	Accessory	glasses, make-up, piercings
	Expression	happiness, surprise, fear
	Color	skin tone, hair color
Text-based	Unchangeable	shape of letters, security features and layout of checks
	Changeable	font type, letter thickness, amount of money and recipient name on a check
	Color	Color of letters and checks

TABLE II: List of features and their examples included in *Feature* questions.

For the *Similar technology* question, we similarly specified the correct answers with check marks (Fig. 7) and collapsed the 5-point Likert scale to 3 for clarity.

Scoring Predict-Reason-Fix questions: We calculate a single 0-1 score for each image the participant sees (10 total — 2 sets of 5). A correct "not fail" prediction scores a 1. For the "fail" images we use a modified multi-class F1 score for the *Reason* and *Fix* portions, where only the primary reason should be selected. We do not penalize for selecting correct secondary reasons. However, we do penalize for selecting incorrect ones. *Fix*: we assign a score of 1 (correct) if the fix matches a selected failure and the reason for failure is correct. We assign 0.5 (semi-correct) if the fix matches the selected failure, but the failure is incorrect. Otherwise, we assign 0 (incorrect).

After calculating the above scores for each image we average them (see Fig. 5).

IV. Results

Recall that all participants only saw questions they marked as ones they were familiar with. Our *Knowledge source* question (Fig. 2) confirms that most of the participants' understanding of the app came from interaction with it — i.e., experiential learning.

Next, we present results for our three research questions, which explore how this experiential knowledge enables people to reason about: the features (within an image) these apps use (RQ.1: Image Features), their capabilities and limitations (RQ.2: Capabilities and Limitations), and recognize the technology in robots (RQ.3: Similar Technology).

A. RQ 1: Image Features

In **Face-based** apps, participants reported almost the same level of importance for various image features, except for the "Unchangeable" ones (Fig. 3). This feature is reported to be significantly more important in *Lock* than in *Filter*. This indicates that participants have the awareness that *Lock* apps need to recognize a person; and for that, they are required to use/process "Unchangeable" features of the person's face.

In **Text-based** apps, "Color" was reported to be the least important feature for both *Text* and *Check* (Fig. 4). This again shows that participants correctly understand that the color of checks or texts would not impact these apps' performance in performing their tasks.

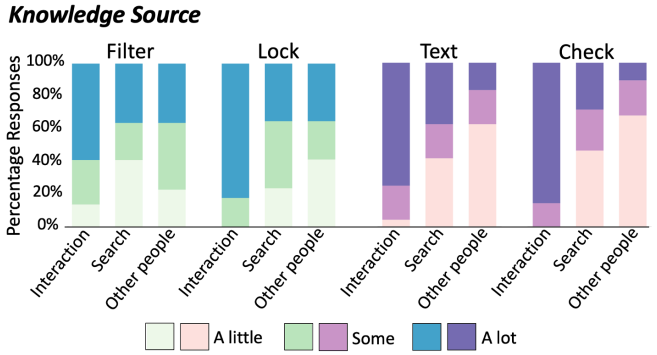

Fig. 2: Source of participants' understanding of how the apps work. Participants gained most of their understanding through interaction with the app.

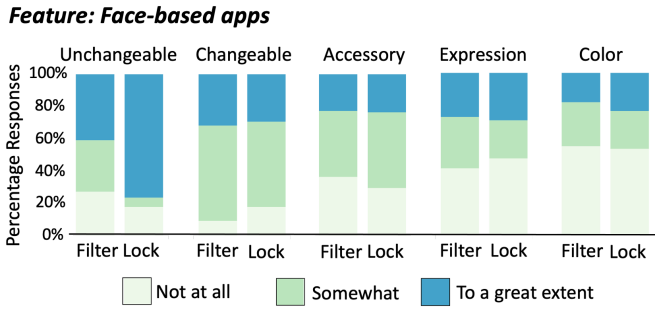

Fig. 3: Summary of the importance of image features for *Face-based* apps reported by participants. *Filter* and *Lock* were similar except for "Unchangeable". "Unchangeable" is reported to be significantly more important for *Lock* than for *Filter*.

B. RQ 2: Capabilities and Limitations

We separate these results into 1) how well the participants could predict the app failure and identify the cause(s) of that failure and 2) what types of capabilities participants assumed these algorithms have.

Predict-Reason-Fix: Overall, participants scored high for all four apps (*Filter*: 0.878, *Lock*: 0.858, *Text*: 0.828, *Check*: 0.841) (see Fig. 5). This indicates that participants were aware both of when the apps fail and what image characteristics caused those failures.

Capability: Participants' responses mostly agreed with our expected answers, indicating that participants' understanding of the capabilities and limitations of the apps are mostly correct. However, there are a couple of notable exceptions (refer to Fig. 6).

First, participants largely disagreed with the "Know" statement, revealing their inclination to use anthropomorphic terms to describe the capabilities of the apps. This was especially true for the **Face-based** ones.

The second exception is concerned with "Distinguish" and "Recognize" statements for *Filter*. These statements expose participants' misconceptions that *Filter* apps can differentiate a specific face from others and recognize a person based on their facial features.

CATEGORY	STATEMENTS	FILTER	LOCK	TEXT	CHECK
Drawing	These apps can distinguish between a real [X] and a cartoon or drawing of one.	A	A	A	A
Know	These apps do not "know" if there is a [X] there - they just look for anything "[X]-like".	A	A	A	A
Distinguish	These apps can distinguish a specific [X] from other [X]s.	D	A	A	A
Recognize	These apps can recognize a specific [X] and know what it looks like.	D	A	A	A
Number	These apps can identify the number of [X]s in the frame.	A	A	A	A

TABLE III: Statements used in the *Capability* question to evaluate participants' understanding of the four apps. The correct answers are specified with A (i.e., Agree) and D (i.e., Disagree). ([X] = face for *Filter* and *Lock*, [X] = text and check for *Text* and *Check*, respectively.)

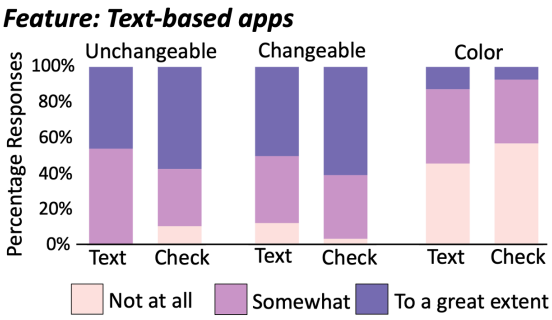

Fig. 4: Summary of the importance of image features for *Text-based* apps reported by participants. "Color" is reported to be of least importance for both apps.

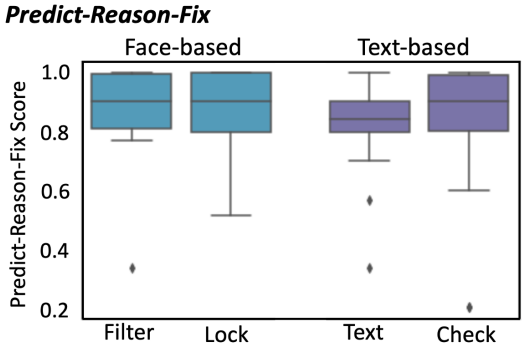

Fig. 5: Predict-Reason-Fix scores for Filter, Lock, Text, and Check apps (2 Predict-Reason-Fix questions for each application type). Overall, participants received high scores for all four apps.

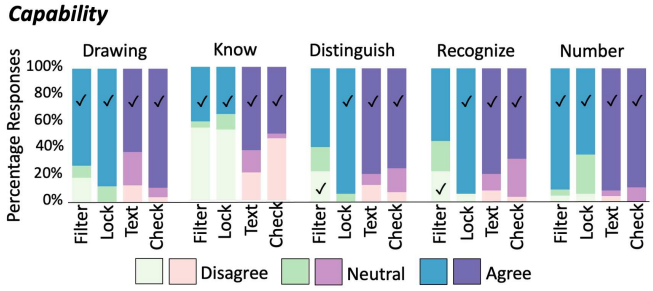

Fig. 6: Summary of participants' level of agreement to the statements in the *Capability* question. Check marks indicate the expected answers. Participants' answers were mostly in agreement with the expected answer except for the "Distinguish" and "Recognize" statements for *Filter* apps and the "Know" statement.

APP	NO. OF PARTICIPANTS	GENDER	AGE	FAMILIAR WITH CV?
Filter	22	Men: 11 Women: 11	18-30: 7 30-50: 10 50+: 5	Yes: 3 No: 19
Lock	17	Men: 8 Women: 9	18-30: 6 30-50: 8 50+: 3	Yes: 2 No: 15
Text	24	Men: 15 Women: 9	18-30: 7 30-50: 9 50+: 8	Yes: 8 No: 16
Check	28	Men: 20 Women: 8	18-30: 9 30-50: 8 50+: 11	Yes: 9 No: 19

TABLE IV: Number of participants per application type and the corresponding demographic information.

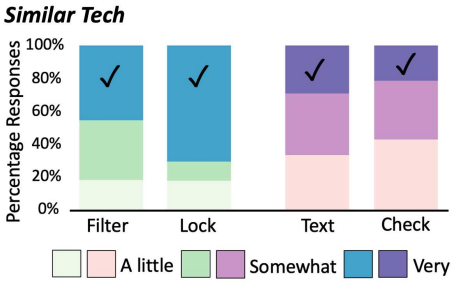

Fig. 7: Summary of the extent to which participants believed a type of robot uses a similar technology to the app. Check marks indicate the expected answers. Only the *Lock* technology was associated with a robotic application.

C. RQ 3: Similar Technology

Participants mostly did not associate the computer vision technology in the app with a similar robotic application (Fig. 7). The strongest association was with the *Lock* application, perhaps because of its use in identifying a patient in a hospital (refer to Table V).

V. Discussion

Our overarching goal is to determine if experiential knowledge gained from use of everyday software apps that utilize computer vision might enable the public to reason about the same technology when embodied in robotic systems.

Our results show that participants reliably predicted the function, failures, and capabilities of software applications incorporating computer vision that they use routinely. Additionally, participants effectively identified image features central to the function of these algorithms. Unfortunately, our

APP	SIMILAR ROBOT TECHNOLOGY	CORRECT ANSWER
Filter	Receptionist robots that greet guests at hotels once they detect a person is looking at them.	Very similar
Lock	Hospital assistant robots that deliver medicines to patients (for patient identification).	Very similar
Text	Room service robots that scan wayfinding signs to navigate in a hotel.	Very similar
Check	Hospital assistant robots that deliver medicines to rooms (for room number identification).	Very similar

TABLE V: List of similar robot technology for each application type and their expected answer.

findings also show that participants do not naturally associate this everyday technology with related computer vision tasks a robot needs to do. Our findings also show that participants ascribed "knowing" to these apps, reaffirming the tendency of people to anthropomorphize this type of technology.

An interesting follow-up research direction would be to more directly ask participants to clarify what the algorithm "knows" and link what it "knows" to observed capabilities/failures.

We propose that experiential knowledge gained through interactive experiences with everyday software applications is an effective way to "calibrate" user expectations of the function and failure states of complex systems. An open question is how to help people link their experience(s) with these common apps to components of more complicated systems, such as a robot.

Our work has implications for AI educators, AI application designers, robot designers, scientists in the field of XAI, and explainable agency. Our results provide insights for AI educators to improve the efficacy of their training approaches and accelerate the learning process by incorporating learners' pre-existing knowledge about the capabilities and limitations of AI apps developed through interacting with them. Moreover, AI educators can highlight and address the misconceptions that learners have developed over the years when using these apps. For instance, this study uncovered two misconceptions specific to *Filter* apps (i.e., *Filter* apps can distinguish a specific face from other faces; *Filter* apps can recognize a person based on their facial features). Identifying and addressing such misconceptions would allow learners to unlearn the misconceptions and learn the concepts that the training material intends to teach more effectively. Designers could also use the findings of this study to improve users' interaction with AI apps and robots by designing various cues in AI apps (e.g., explanatory texts, visual stimuli) and robots (e.g., verbal and non-verbal) to help the users understand the actual capabilities and limitations of such devices and correct their misconceptions. For scientists in the fields of XAI and explainable agency, this study suggests considering the experiential knowledge of users and their misconceptions, especially AI novices, of AI daily apps and avoiding the use of anthropomorphic terms such as "Know" in the design of explanations to prevent the formation of inaccurate expectation of AI systems.

VI. Acknowledgment

We thank Mahsa Saeidi for her insights. Funded in part by NSF grants NRI 2024872, 2024673 and 2024643.

References

[1] 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.

[2] T. Williams, P. Briggs, and M. Scheutz, "Covert robot-robot communication: Human perceptions and implications for human-robot interaction," *Journal of Human-Robot Interaction*, vol. 4, p. 23, 05 2015.

[3] M. de Graaf, S. Allouch, and J. A. Van Dijk, "Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study," 03 2017.

[4] J. A. Fails and D. R. Olsen, "Interactive machine learning," in *Proceedings of the 8th International Conference on Intelligent User Interfaces*, ser. IUI '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 39–45.

[5] R. Fiebrink, P. R. Cook, and D. Trueman, "Human model evaluation in interactive supervised learning," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 147–156.

[6] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ser. IUI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 126–137.

[7] W. Rouse and N. Morris, "On looking into the black box. prospects and limits in the search for mental models," *Psychological Bulletin*, vol. 100, 10 1984.

[8] 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.

[9] J. R. Williford, B. B. May, and J. Byrne, "Explainable face recognition," *CoRR*, vol. abs/2008.00916, 2020.

[10] B. R. Webster, S. Y. Kwon, C. Clarizio, S. E. Anthony, and W. J. Scheirer, "Visual psychophysics for making face recognition algorithms more explainable," *European Conference on Computer Vision 2018*, vol. 15.

[11] 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*. Curran Associates, Inc., 2016, pp. 2280–2288.

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

[13] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *CoRR*, vol. abs/1706.07269, 2017.