Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application

SUNNIE S. Y. KIM, Department of Computer Science, Princeton University, USA
ELIZABETH ANNE WATKINS*, Intelligent Systems Research, Intel Labs, USA
OLGA RUSSAKOVSKY, Department of Computer Science, Princeton University, USA
RUTH FONG, Department of Computer Science, Princeton University, USA
ANDRÉS MONROY-HERNÁNDEZ, Department of Computer Science, Princeton University, USA

Trust is an important factor in people's interactions with AI systems. However, there is a lack of empirical studies examining how end-users trust AI in real-world contexts. Most research investigates one aspect of trust in lab settings with hypothetical end-users. In this paper, we provide a holistic and nuanced understanding of trust in AI through a qualitative case study of a real-world computer vision AI application. We report findings from interviews with 20 end-users of an app for bird identification where we inquired about their trust in the app from many angles. We find participants perceived the app as trustworthy and trusted it, but selectively accepted app outputs only after engaging in verification behaviors, and decided against app adoption in certain high-stakes scenarios. We also find domain knowledge and context are important and influential factors of participants' trust-based cognitive assessments and behavioral decision-making. We discuss the implications of our findings and provide recommendations for future research.
** This work is a shorter version of [16]. The full paper is available at https://sunniesuhyoung.github.io/XAI_Trust.*

1 INTRODUCTION

Trust is a key factor in people's interactions with Artificial Intelligence (AI) systems. For the effective adoption and use of these systems, people must trust them appropriately. Both unwarranted trust (trusting when the AI system is not trustworthy) and unwarranted distrust (distrusting when the AI system is trustworthy) can hurt the quality of interactions [10, 22, 32]. To better understand trust and foster it appropriately in human-AI interactions, recent works have started to investigate questions such as: What does it mean to trust an AI system? [10] How is trust established and developed? [20] What factors influence people's trust and how? [33, 36]

Trust in AI research, however, is still in a nascent stage. As noted in recent surveys [28, 29], papers often use different definitions of trust, making their results difficult to compare. There is also little agreement on how to empirically study trust, e.g., when to use subjective vs. objective measures. Finally, there is a lack of research that approaches trust holistically. Most papers study one specific aspect of trust (e.g., whether explainable AI increases people's trust in AI systems [33, 36]) in artificial lab settings with hypothetical end-users. While they provide valuable insights, they do not capture the complex nuances of trust in real-world contexts.

In this work, we provide a more holistic and nuanced understanding of trust in AI through a qualitative case study of a real-world AI application. We ground our study in Merlin, a mobile phone app that uses computer vision AI models to identify birds in user-uploaded photos and audio recordings. Concretely, we conducted semi-structured interviews with 20 Merlin end-users and inquired about their trust in the app from several angles. Since we are one of the first to talk to actual end-users about their trust in the AI application, we focus on exploring and understanding different aspects of trust and the factors that influence them, rather than quantifying the importance of certain pre-specified factors.

We make three key contributions in this work: (1) We study how end-users trust AI in a real-world context and what factors influence their trust. In doing so, we synthesize theoretical and empirical research by applying theoretical

* Most work done while at Princeton Center for Information Technology Policy and Human-Computer Interaction Program.

definitions and models of trust to empirical data from the real world. This approach will, we hope, yield insights into how readily these theoretical models can be operationalized for empirical research. (2) We provide a more nuanced understanding of trust, as compared to the current state of the art in the field. We find general trust attitudes and trustworthiness perceptions are distinct from trust-related behaviors. While our participants told us they perceived the AI system as trustworthy and trusted it, they also described how they selectively accepted AI outputs only after engaging in verification behaviors, and sometimes decided against AI adoption in certain high-stakes scenarios. Key for our pursuit of how trust-based user assessments happen in real-world settings, we also find that end-users' domain knowledge, and context of use, to be important factors of trust-based decision-making. (3) Finally, we discuss the implications of our findings and provide recommendations for future research. Most critically, we advocate for researchers to define and delineate trust from related constructs, and to consider human, AI, and context-related factors of trust together.

2 BACKGROUND AND RELATED WORK

2.1 Definitions and models of trust in AI

According to recent surveys of trust in AI research [8, 28, 29], many works do not state a definition of trust. Among works that do, the most commonly used definitions are: (1) "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party" by Mayer et al. [21] and (2) "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" by Lee and See [19].

Both definitions share the same key elements, as described by Vereschak et al. [29]: (1) *vulnerability*: the situation involves uncertainty of outcomes and potential negative consequences; (2) *positive expectations*: the trustor thinks that negative outcomes associated with trusting do not exist or are very unlikely; (3) *attitude*: the general way of thinking and feeling, typically reflected in a behavior, although not a behavior itself. These elements also distinguish trust from other related constructs. It is not *trust* but: *confidence* when there is no vulnerability; *distrust* when there is no positive expectation; *compliance* or *reliance* when referring to a behavior; and *perceived trustworthiness* when referring to a perception of a trustee's characteristics upon which trustors form their trust.

Recently, scholars have proposed specific definitions and models for "trust in AI" [10, 20]. Jacovi et al. [10] formalized trust in AI as "contractual": to trust an AI system is to believe that it is trustworthy to uphold some contract. Their formalization disentangles trust and trustworthiness, and defines "warranted trust" as trust that is "caused" by the AI's trustworthiness. Liao and Sundar [20], on the other hand, took a communication perspective and proposed a model that describes how the trustworthiness of AI systems is communicated through trustworthiness cues and how those cues are processed by people to make trust judgments.

In this work, we adopt the model of trust by Mayer et al. [21] because the model's definition and process orientation fit our work's objective of understanding and disentangling different aspects of trust in AI. Mayer et al. [21] delineate trust from its antecedents, context, and products, and describe how different components influence each other as the trustor interacts with the trustee (Fig. 1). Based on their model, we separate trust from trustworthiness perceptions (trustworthiness being trust's antecedent) and trust-related behaviors. The models used by Jacovi et al. and Liao and Sundar [10, 20] are less fitting because Jacovi et al. [10] focus on formalizing prerequisites, causes, and goals of trust in AI, and Liao and Sundar [20] focus on modeling the communication of trustworthiness.

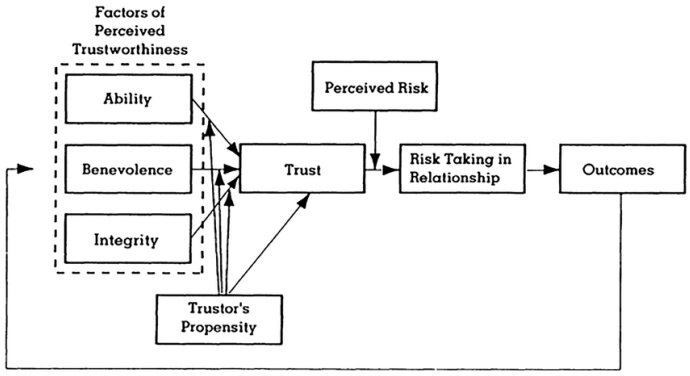

Fig. 1. Based on Mayer et al.'s trust model [21], we separate *trust* from *trustworthiness* perceptions that precede it and trust-related behaviors that proceed from it: (1) *risk taking in relationship* which we call AI *use* or *adoption*, and (2) *outcomes* evaluation which we call AI *output acceptance* throughout the paper.

2.2 Empirical studies of trust in AI

Much of prior work focused on understanding the effect of certain *pre-specified factors* on trust [1, 2, 5, 14, 17, 18, 23–26, 30, 33, 34, 36]. Most utilized *lab experiments*, usually with participants recruited from crowdsourcing platforms (e.g., MTurk [2, 5, 14, 18, 23–26, 33, 36], Prolific [23, 24], internal platform [17]). Typically, these works start with a hypothesis (e.g., explainability will increase trust in AI). To investigate the hypothesis, they choose a measure of trust (e.g., self-reported rating on a 1-7 scale [12]), make a change to the factor of interest in the design of the AI system (e.g., show an explanation of the AI's output), and then quantify the effect of that change on participants' trust. Based on the results, they conclude the effect of the factor of interest on trust.

The most commonly studied factors in the literature are transparency and explainability. However, researchers operationalize these factors in several ways. For instance, *transparency* is operationalized as providing model internals (e.g., learned coefficients in a linear regression model) in [25], overall performance measures (e.g., accuracy) in [17, 18, 26, 33, 34], confidence scores of individual outputs in [23, 36], and visualizations of input data distributions and feature engineering process in [6]. Similarly, while [2, 4, 5, 14, 17, 18, 23, 24, 26, 33, 35] all study the effect of *explainability* on trust, the operationalized explanations of AI's behavior and outputs greatly vary in approach (e.g., feature attribution, counterfactual examples) and form (e.g., heatmap-based, part-based).

These works provide insights into the relationship between trust and the factor of interest, as operationalized in a specific and controlled way. However, they do not capture the contextual aspects of trust in real-world human-AI interactions, and the design of these studies does not allow for discovering new trust-influencing factors. To address these two gaps, we conducted a qualitative case study of a computer vision app and interviewed its end-users about their trust in it. While resource-intensive, interviews enabled us to explore multiple aspects of trust in depth and identify old and new factors of trust in AI. The value of qualitative case studies has been demonstrated in recent works [7, 27, 31]. In one example, Widder et al. [31] conducted a case study investigating what factors influence engineers' trust in an autonomous software engineering tool in a high-stakes workspace. They found that trust, in their study setting, was influenced by the tool's transparency, usability, social context, and the organization's associated processes. Widder et al.'s work lays groundwork for our own qualitative study, as we apply their methods to asking similar questions about a different population working in a different domain: the end-users of a computer-vision app.

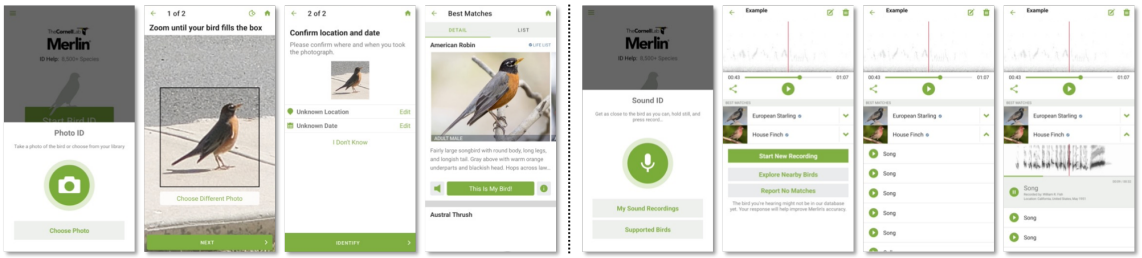

Fig. 2. Merlin is an AI-based bird identification mobile phone app. Users upload photos on the *Photo ID* feature (left) or sounds on the *Sound ID* feature (right), with optional location and season data, to get a list of birds that best match the input.

3 METHODS

In this section, we describe our study methods which were reviewed and approved by our Institutional Review Board.

To study trust in AI in a realistic setting, we looked for a research scenario that first, involves real-world AI use by end-users who range in their domain and AI knowledge base, and second, satisfies the requirements of widely-accepted trust definitions [19, 21]. We found the Merlin app (Fig. 2) to satisfy both conditions. First, Merlin is a mobile phone app that identifies bird species from user-input photos and/or audio recordings. It is an *expert* application with expertise that most people do not have, i.e., knowledge and skill to identify thousands of birds. As a free app with over a million downloads, it is used by people with diverse domain (bird) and AI backgrounds, thus satisfying our first requirement. Second, as we verify in Appendix C, there are *vulnerability* and *positive expectations* in its use, allowing us to characterize the app end-users' *attitude* toward the app as trust, and study "trust in AI" and their influencing factors.

We recruited 20 participants who are active end-users of Merlin Photo ID and/or Sound ID, the app's AI-based bird identification features, with considerations for diversity in the domain and AI background (see Appendix A for details). From July to August 2022, we interviewed participants over a Zoom video call and inquired about their trust in the app (see Appendix B for the interview protocol).[1] In summary, we asked about general attitudes and perceptions, such as how accurate and trustworthy they find the app, as well as specific instances, such as how they assess the correctness of the app's outputs, and in what circumstances they decide to use the app and not. We also asked whether they would adopt the app in two hypothetical high-stakes scenarios with health-related and financial outcomes:

(1) *Sick bird scenario*: Suppose you find a sick bird and take it to the vet. The vet is not sure what bird it is. Would you recommend Merlin to identify the bird species so that the vet can determine the course of treatment?
(2) *Game show scenario*: Suppose you enter a game show where you can win or lose money based on how well you can identify birds from photos or audio recordings. You can only use one resource among Merlin, books (e.g., field guides), the Internet (e.g., search engine, online birder community), and so on. Which resource would you use? Does your answer change depending on certain factors?

We transcribed the interviews and then analyzed the transcripts using descriptive coding. We first read five transcripts to develop an initial codebook, then collectively iterated on and refined the codebook over multiple meetings. After all data were coded, we discussed the results and drew out themes.

[1] In the same interviews, we also inquired about participants' explainability needs, intended uses of AI explanations, and perceptions of existing explanation approaches, and analyzed that portion of the data in another paper [15].

Table 1. Factors that influenced participants' trust in AI.

Human-related factors	AI-related factors	Context-related factors
Domain knowledge	Ability	Task difficulty
Ability to assess the AI's outputs	Integrity	Perceived risks and benefits
Ability to assess the AI's ability	Benevolence	Situational characteristics
Ability to use the AI	Popularity	Domain's reputation
	Familiarity	Developers' reputation
	Ease of use	

4 RESULTS

We unpack participants' trust in AI in three parts: trustworthiness perception (Sec. 4.1); acceptance of individual AI outputs (Sec. 4.2); and AI adoption (Sec. 4.3). In Tab. 1, we summarize the factors that influenced participants' trust based on whether they are related to the human trustor, the AI trustee, or the context.

4.1 Trustworthiness perception: Participants assessed the AI to be trustworthy

We begin by explaining how participants assessed the app's trustworthiness, a key antecedent to trust in Mayer et al.'s model [21]. Overall, participants assessed that the app possesses ability, integrity, and benevolence–the three factors of perceived trustworthiness in the model (Fig. 1)—and assessed the app to be trustworthy.

4.1.1 Participants assessed the AI's ability based on their prior experience with the AI and the AI's popularity. **Ability** refers to the trustee's skills and competencies [21]. For automation systems, Lee and See [19] describe it as **performance**, i.e., how well the automation is performing. Participants were overall impressed with the app and judged it to have high ability. Most described the app as very successful and that it seemed to be correct 9-10 out of 10 times, based on their **prior experience** with the app. The only participants who did not describe the app as very successful were P2, who said Sound ID often made mistakes, and P6, who was disappointed with Photo ID. Most other participants were impressed and described the app as "*pretty insane*" (P15), "*perfect*" (P11), and gave high praise despite having observed mistakes: "*I love it. I trust it. [...] I've had one or two times where I've thought I don't believe that's really that bird? [...] But I don't deeply care. Doesn't matter. I trust it. I trust it*" (P14). Intriguingly, some participants mentioned they could not accurately assess the app's ability due to their lack of **domain knowledge** (P11, P12, P13). For instance, P11 said: "*As far as I know, it's been perfect, but I don't know enough to know if it would be making mistakes.*" Finally, while most judged the app's ability based on their own prior experience, P12 made an assessment based on the app's **popularity** which is an external factor: "*I imagine that if it has such a wide user base, it would be pretty accurate*" (P12).

4.1.2 Participants assessed the AI's integrity based on developers' reputation. **Integrity** refers to the degree to which the trustee adheres to a set of principles that are acceptable to the trustor [21]. For automation systems, [19] describe it as **process**, i.e., in what manner and with which algorithms it is accomplishing its objective. We found that participants believed the app's integrity because due to the **reputation of the developers**, the Cornell Lab of Ornithology which is a respected institution with a long history of science and conservation efforts. Most participants were well aware that the app was developed by the Cornell Lab of Ornithology (P1, P3, P4, P5, P8, P9, P11, P12, P13, P14, P15, P17, P18, P19). Participants were also familiar with the lab's other apps (e.g., eBird, BirdNET, iNaturalist) and resources (e.g., All About Birds, Macaulay Library), describing these and the app as their "*go-to*" (P15) when they want to learn about a specific

bird. P14 specifically said they trusted the app because it was developed by the Cornell Lab of Ornithology: "*I know that Cornell Ornithology Lab does excellent, excellent stuff. I mean, if you're going to try and learn anything about a bird, just go there. [...] So I trusted it [the app] for that reason.*" Participants did not know how the app was developed or how it works, since such information is not publicly available. Nonetheless, they believed the app's integrity because they believed in the authority and expertise of the app developers.

4.1.3 Participants assessed the AI's benevolence based on the domain's reputation. **Benevolence** refers to the extent to which the trustee's motivations are aligned with the trustors' [21]. For automation systems, Lee and See [19] describe it as *purpose*, i.e., why the automation was built originally. We found that participants believed the app's benevolence because of the positive **reputation of the domain**, i.e., the birding community that they themselves and the app developers are part of. For instance, P18 described the birding community as a place where everyone tries to be accurate and do good: "*I think birders, in general, are a community where there's very few people who try and do adversarial attacks because it doesn't benefit anybody [...] the value of the birding community is that everybody is trying to be accurate.*" Some participants contrasted the app with other AI applications. For instance, P2 described the app as not having "*malicious intent*" compared to advertisements. P9 contrasted the app with other AI applications they found "*creepy*" and "*harmful,*" such as voice assistants that may be "*monitoring*" user behavior.

4.2 Output acceptance: Participants selectively accepted AI outputs only after verification

Participants described the app as "trustworthy"; however, they did not accept its outputs as "true" in every single instance of use. To the extent possible, participants carefully evaluated outputs and made acceptance decisions after verification. Participants' ability to assess the app outputs, however, heavily depended on their knowledge.

4.2.1 Participants verified AI outputs using their domain knowledge. Participants assessed the app's outputs using various verification methods. One method was assessing the **likelihood** of spotting a bird species in a given area (P1, P2, P6, P16, P17, P19). Participants described they were more trusting of the app's output when the identified species is common for the area, and less trusting when it is rare. For instance, P6 said: "*If it's a common bird or even just a rare bird, uncommon or something like that, then maybe [it is correct]. But if it's a super rare bird, then definitely not.*"

Participants also assessed **task difficulty**. P1's response well explains the relevant reasoning: "*I trust it [the app] more when I know that I'm looking at something that should be relatively unambiguous. If I'm looking at something that's like a Female Warbler or a Female Sparrow, which might just be a little brown bird, then I'm a little bit more skeptical of the result.*" For context, "*little brown bird*" is a term used by birders to describe a large number of species of small brown passerine birds, which are known to be notoriously difficult to distinguish. P1 described them as "*really hard to ID, even for a human ornithologist.*" Note that assessment of likelihood and task difficulty requires **domain knowledge** of which birds are common and rare for the area, and which birds are difficult and easy to identify.

Several participants compared input photos and audio recordings to reference photos and audio recordings of the identified species, which are provided in the app (P1, P10, P20). This verification does not require **domain knowledge** per se; however, participants with it could more easily verify the output as they would know what to check. Some participants used information from other sources (P1, P4, P10, P15, P18). If the app identified a bird based on sound, P10 and P18 tried to confirm it with their own visual identification, and vice versa. P1, P4, and P15 took a step further and consulted other birders, through their personal networks or online communities. We note that cross-checking requires **domain knowledge** for identifying birds on their own, whereas consulting other birders does not.

4.2.2 Some participants disregarded AI outputs when they could not verify. For some participants, verification was a crucial and necessary step for output acceptance (P3, P4, P15, P18). When unable to verify, they disregarded the app's output. For example, P15 said they've never only relied on the app when identifying a bird they have not seen before. They almost always sent the app output to more experienced birders and received their confirmation. P18 was also strict about when they accept the app outputs, stating, "*I never, I never count on my bird registry anything that Sound ID says that I can't kind of confirm either through the facts of it or through a visual ID*" (P18). These participants disregarded unverifiable app outputs, despite their positive assessment of the app's ability and trustworthiness, revealing a gap between general trustworthiness perceptions and instance-specific trust-related behaviors.

4.2.3 Not all participants had the ability to assess the correctness of AI outputs. So far we described various processes through which many participants decided whether or not to accept app outputs. However, not all participants had the **ability to assess** the correctness of app outputs. In Sec. 4.1, we described how some participants with little **domain knowledge** said they could not accurately assess the app's ability (P11, P12, P13). These participants also said that because they "*know so little about birds*" (P12), they could not "*validate or reject*" (P11) the app outputs, especially if they can't get information from other sources. P13 said, "*If it's misidentifying a bird that I can't see, then I have no way to know that.*" This finding suggests that domain knowledge has a wide influence on participants' interactions with the app.

4.3 Adoption decision: Participants never decided against using the AI in their actual use setting, but carefully made AI adoption decisions for hypothetical high-stakes scenarios

Finally, we describe how participants made AI adoption decisions. To gain a richer understanding of this decision-making process, we compared participants' decisions between their actual use setting and two hypothetical high-stakes scenarios (see Appendix B for details). We found that while participants always used the app in their actual use setting, they made different adoption decisions for the high-stakes scenarios based on various factors: the app's ability, familiarity, and ease of use; participants' ability to assess the app's outputs and use the app; and finally, task difficulty, perceived risks and benefits of the situation, and other situational characteristics.

4.3.1 In their actual use setting, participants never decided against using the AI. We found that participants always use the app when opportunities arise. It is not that participants absent-mindedly used the app. Participants were aware of when the app works well and not, and knew how to help the app be more successful, e.g., by supplying better inputs. However, when we asked how they make app adoption decisions, participants only described situations where they decided to use the app, and never situations where they decided against using it.

There could be several reasons for this finding. First, the app has a low cost of use. The only costs are the time and effort involved in taking photos or audio recordings and inputting them into the app, and perhaps a small amount of phone battery. Second, the risks of use are also low. There are potential negative consequences when the app misidentifies, e.g., gaining wrong knowledge, as described in **??**. However, end-users can mitigate these risks by verifying the output and rejecting it if needed. Finally, we only interviewed active end-users of the app, who are likely to continue to use the app because they are satisfied with it. Past or non-users may provide different responses.

4.3.2 In hypothetical high-risk scenarios, participants carefully considered the AI's ability and various contextual factors. When we presented participants with hypothetical high-risk scenarios, we observed a different decision-making process around app adoption. Participants considered the app's **ability** with respect to various **situational characteristics**. For example, for the sick bird scenario, some participants judged the app is worth a try. P15 described using the app as

"*something that wouldn't hurt*" since they are in a situation where both they and the vet could not identify the bird. They expressed some degree of confidence in the app's ability: "*I feel like Merlin's not gonna tell you that a baby hawk is a chickadee*" (P15). Similarly, P17 said they "*would definitely recommend it [the app] to get into the right ballpark.*" Still, they recommended consulting other scientific resources and doing a "*triple check*" of the app's output, since the **risk** of misidentification, i.e., the sick bird receiving the wrong treatment, is higher than the risk in their actual use settings, e.g., gaining wrong knowledge.

Other participants were skeptical that the app could identify the sick bird (P3, P6, P15, P19). P4 did not think the app could identify birds that they and the vet could not: "*Assuming that I don't know what the bird is and they [vet] don't know what the bird is, this bird is some ambiguous-looking bird. In those cases [...] I don't think Merlin would be able to know.*" P15 pointed out that sick birds are often "*fledglings, juveniles*" which are "*harder to ID for everybody in real life and presumably harder for Merlin.*" P6 noted that sick birds may be "out-of-distribution" for the app due to their underrepresentation in the training data: "*I assume Merlin is not trained on sick birds, so I can totally see it doing something crazy.*" These participants weighted the app's ability against the **task difficulty** and decided against adopting the app in the sick bird scenario.

Similarly, for the game show scenario, participants jointly considered the app's ability and the situation's characteristics. P1 and P4 said they would choose the app if there are **time constraints**, but otherwise choose "*a good quality field guide*" (P1) or "*Discord*" (P4). Similarly, P15 said they would choose the app if they need to give an answer "*quickly, within 30 seconds,*" but otherwise consult other birders. Others said they would choose the app if they have to do **sound-based identification** on the game show (P6, P10, P12, P20), describing difficulties with text-based referencing of sound: "*some books saying 'it goes da-da-da' is not helpful*" (P6). P20 explained their reasoning in detail: "*If I'm on this game show and it plays a sound, I would definitely want to use the app [...] but if it shows me a bird, I might just want to google it because I have enough knowledge personally that I could probably guess what type [...] and then search by colors. So I guess it comes down to what I think the app does the best, which is sound, versus what I think I can get away without it.*" Participants considered the app's ability not only on its own, but also in comparison to other resources. Again, participants carefully made app adoption decisions as the **perceived risks and benefits** of the scenario, i.g., loss and gain of money, are higher than those of their actual use settings.

4.3.3 Some participants adopted the AI due to familiarity and ease of use. Two other factors that drove participants' app adoption decisions were **familiarity** and **ease of use**. For the sick bird scenario, P2 said they would definitely used the app because it "*feels kind of like second nature.*" P16 also chose to identify the sick bird with the app because using the app "*would be the easiest.*" Similarly, for the game show scenario, P2 picked the app as their top choice because "*it's so easy [...] it doesn't take all that much time to look through everything.*" P16 mentioned both familiarity and ease of use: "*I think Merlin would make the most sense since I'm familiar with it.*" They described other resources as requiring more "work" by end-users, compared to the app where end-users can just input bird photos and/or audio recordings: "*You still have to do a lot of work to do like a Google search compared to this [app]*" (P16).

P4 described another aspect of familiarity: their **ability to use the app**. They said, "*I definitely would use Merlin because I'm familiar with it. And I trust my ability, like I know how to operate it pretty well.*" (P4). We found this response particularly interesting because the ability to use the AI has not been explored much in the trust in AI literature. However, we expect it will become an important topic in trust and human-AI interaction research, as AI applications grow in complexity and require end-users to develop skills for effective AI use.

5 DISCUSSION AND CONCLUSION

5.1 Key findings and their implications

5.1.1 The complex picture of trust. The key takeaway from our study is that end-users' trust relationship with AI is complex. Participants found the app overall trustworthy, but still carefully assessed the correctness of individiual outputs and decided against adoption in certain high-stakes scenarios. This finding illustrates that trust is a multifaceted construct that must be approached holistically. It further urges researchers to study multiple aspects of trust together. To get a full and accurate picture of trust, it is crucial to examine both *general* aspects such as trustworthiness perceptions and *specific* aspects such as AI output acceptance and adoption decisions.

5.1.2 Insights from specific trust-related behaviors and the importance of domain knowledge. Participants' instance-specific decisions about AI output acceptance and adoption were particularly useful for understanding what factors influence trust in AI and how. For example in Sec. 4.2, we described how participants trusted the app's output more when the task is easy (e.g., "*relatively unambiguous*" bird) and less when the task is difficult (e.g., "*little brown bird*"). Similarly in Sec. 4.3, we described how some participants were hesitant to use the app to identify the sick bird in the hypothetical scenario because they judged the task would be too difficult for the app. These examples illustrate the rich reasoning behind participants' trust-related behaviors, where factors of trust interact with each other. Participants used their domain knowledge (human-related factor) to assess task difficulty (context-related factor) and weighted it against the app's ability (AI-related factor) to decide whether to accept the app's output or adopt the app in the given situation.

In earlier sections, we also described domain knowledge's influence on participants' ability to assess the app's ability (Sec. 4.1) and outputs (Sec. 4.2). Participants with domain knowledge assessed the correctness of the app outputs by, for example, cross-checking it with their own identification and judging its likelihood based on their knowledge of what birds are common and rare in the area. Participants without domain knowledge, however, had difficulties in assessing the app outputs and, consequently, the app's overall ability. Based on these findings, we urge the community to consider domain knowledge when designing AI applications and trust calibration interventions. For example, a system could assess the verification behaviors used by domain experts, and build these options into the system so that they are accessible to experts and non-experts alike.

5.1.3 The importance of contextual factors and contextually-grounded studies. Finally, we highlight the importance of contextual factors and contextually-grounded studies for understanding their influences on trust. When participants were describing the app's trustworthiness, we observed that positive reputation of the domain (birding community) and developers (Cornell Lab of Ornithology) led them to positively assess the AI's ability, integrity, and benevolence (Sec. 4.1). We have two points of discussion on this finding. First, it shows that external contextual factors (reputation of the domain and developers) influence internal AI factors (ability, integrity, and benevolence), underlining the impact of contextual factors on trust. It also reiterates that factors influence each other, and calls for research that studies the interactions between factors. Second, while the specific finding is context-dependent, the provided insight can generalize to other types of AI applications. For example, we can anticipate end-users to have doubts about an AI application's ability if the AI is not developed by a well-known institution; benevolence if the AI seems to have a different goal from them (e.g., recommendation systems trying to sell unneeded products); and integrity if the AI seems to make decisions with wrong reasons (e.g., decision-making systems discriminating based on protected attributes).

5.2 Adapting existing trust models to AI

Overall, we found Mayer et al.'s model [21] useful for understanding trust in AI. In particular, we found helpful the way in which it breaks down "trust" into multiple components and delineates trust from its antecedents, context, and products (Fig. 1). However, as with any model, there were some limitations and challenges in its application. First, since the model was originally developed for *trust between people*, we had to make adaptations to apply it to *trust in AI*. For instance, when describing participants' trustworthiness perceptions (Sec. 4.1), instead of using [21]'s definitions of ability, integrity, and benevolence, we used Lee and See's [19] automation-friendly translations of these factors: performance, process, and purpose. Second, Mayer et al.'s model [21] is by no means a comprehensive trust model. This is expected as the paper's goal was not to list all possible antecedents of trust. Hence, we drew from other works [1, 9, 11, 13] to categorize the factors we identified into human, AI, and context-related factors. Third, we did not observe the influence of trustor's propensity, one of the model components, in our study. However, this result does not imply that trustor's propensity is an unimportant factor of trust in AI. Future research, in particular survey and experimental studies, are needed for such conclusions.

5.3 Limitations and recommendations for future work

Finally, we discuss our work's limitations and provide recommendations for future work. First, as with any qualitative case study, our findings may not generalize to other settings. This is an intentional trade-off made in favor of gaining an in-depth understanding of how end-users trust and interact with AI in a specific context. Another limitation is that all participants were active end-users of the app. Those who stopped using it or chose to not use it are not represented in the study. Finally, due to the highly multifaceted and dynamic nature of trust in AI, there are aspects of it that our work does not cover, e.g., how trust changes overtime and how trust processes vary between applications.

We conclude with a set of practical recommendations for future research on trust in AI. **(1) State a definition of trust.** Trust is a multifaceted construct that carries different meanings to different people. Explicitly stating a definition of trust can help remove ambiguity and confusion around the term, and help researchers accurately interpret and compare study results. **(2) Examine if trust is the construct being studied.** Depending on the study design and context, what's being studied may not be trust at all, but other related constructs such as confidence and reliance. We recommend that researchers carefully examine their study design and context to ensure trust is the construct being analyzed. We hope our Appendix C serves as a helpful example of such an examination. **(3) Approach trust holistically and study its antecedents, context, and products.** Our contextually-grounded study of trust, trustworthiness perceptions, and trust-related behaviors revealed a comprehensive picture of end-users' trust relationships with AI that cannot be gained by studying only trust. Hence, we recommend studying trust together with its antecedents, context, and products, to the extent possible. **(4) Consider human, AI, and context-related factors and their interactions.** As observed in this work, trust in AI is influenced by many factors. To prevent surprises and gain a comprehensive understanding of trust in a given context, we recommend anticipating as many factors as possible and studying their interactions. We found it particular helpful to consider factors along the dimensions of human, AI, and context. We hope our work aids future research on other AI applications and the disparate contexts into which they are integrated.

ACKNOWLEDGMENTS

We foremost thank our participants for generously sharing their time and experiences. We also thank Tristen Godfrey, Dyanne Ahn, and Klea Tryfoni for their help in interview transcription. Finally, we thank Angelina Wang, Vikram V.

Ramaswamy, Amna Liaqat, Fannie Liu, and other members of the Princeton HCI Lab and the Princeton Visual AI Lab for their helpful and thoughtful feedback. This material is based upon work partially supported by the National Science Foundation (NSF) under Grants No. 1763642 and 2145198 awarded to OR. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We also acknowledge support from the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), Princeton SEAS Project X Fund (RF, OR), Princeton Center for Information Technology Policy (EW), Open Philanthropy (RF, OR), and NSF Graduate Research Fellowship (SK).

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2022. Hartmann, Philipp and Hobert, Sebastian and Schumann, Matthias. *Americas Conference on Information Systems (AMCIS)* 5 (2022). https://doi.org/amcis2022/sig_ed/sig_ed/5

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[3] Michaela Benk, Suzanne Tolmeijer, Florian von Wangenheim, and Andrea Ferrario. 2022. The Value of Measuring Trust in AI - A Socio-Technical System Perspective. https://doi.org/10.48550/ARXIV.2204.13480

[4] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[5] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[6] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 297–307. https://doi.org/10.1145/3377325.3377501

[7] Madeleine Clare Elish and Elizabeth Anne Watkins. 2020. Repairing innovation: A study of integrating AI in clinical care. *Data & Society* (2020).

[8] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (2020), 627–660. https://doi.org/10.5465/annals.2018.0057 arXiv:https://doi.org/10.5465/annals.2018.0057

[9] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570 arXiv:https://doi.org/10.1177/0018720814547570 PMID: 25875432.

[10] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[11] E Jermutus, D Kneale, J Thomas, and S Michie. 2022. Influences on User Trust in Healthcare Artificial Intelligence: A Systematic Review [version 1; peer review: 1 approved with reservations]. *Wellcome Open Research* 7, 65 (2022). https://doi.org/10.12688/wellcomeopenres.17550.1

[12] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04 arXiv:https://doi.org/10.1207/S15327566IJCE0401$_0$4

[13] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. 0. Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors* 0, 0 (0), 00187208211013988. https://doi.org/10.1177/00187208211013988 arXiv:https://doi.org/10.1177/00187208211013988 PMID: 34048287.

[14] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the Human Interpretability of Visual Explanations. In *European Conference on Computer Vision (ECCV)*.

[15] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

[16] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[17] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[18] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for

Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[19] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 arXiv:https://doi.org/10.1518/hfes.46.1.50$_3$0392 PMID: 15151155.

[20] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182

[21] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792

[22] Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research? https://doi.org/10.48550/ARXIV.2209.00651

[23] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Neural Information Processing Systems (NeurIPS)*.

[24] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. In *Neural Information Processing Systems (NeurIPS)*.

[25] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[26] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *IUI*.

[27] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. https://doi.org/10.1145/3351095.3372827

[28] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 254, 7 pages. https://doi.org/10.1145/3491101.3519772

[29] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (oct 2021), 39 pages. https://doi.org/10.1145/3476068

[30] Weiquan Wang and Izak Benbasat. 2008. Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for E-Commerce. *Journal of Management Information Systems* 24, 4 (2008), 249–273. http://www.jstor.org/stable/40398919

[31] David Gray Widder, Laura Dabbish, James D. Herbsleb, Alexandra Holloway, and Scott Davidoff. 2021. Trust in Collaborative Automation in High Stakes Software Engineering Work: A Case Study at NASA. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 184, 13 pages. https://doi.org/10.1145/3411764.3445650

[32] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. [n. d.]. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3581197

[33] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[34] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[35] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2016. Top-down Neural Attention by Excitation Backprop. In *European Conference on Computer Vision (ECCV)*.

[36] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

APPENDIX

A PARTICIPANT RECRUITMENT AND SELECTION

We recruited participants who are active end-users of Merlin Photo ID and/or Sound ID, the app's AI-based bird identification features, with considerations for diversity in the domain and AI background. Concretely, we created a

Table 2. Participants' domain (bird) and AI background.

	Low-AI	**Medium-AI**	**High-AI**
Low-domain	P7, P12, P16	P8, P14	P11, P13
Medium-domain	P2, P20	P1, P4, P10	P6
High-domain	P5, P17	P3, P9, P15	P18, P19

screening survey with questions about the respondent's background and app usage pattern (e.g., regularly used features, frequency of use). We posted the survey on various channels: Birding International Discord, AI for Conservation Slack, several Slack workspaces within our institution, and Twitter. In addition to posting the survey on Twitter, we reached out to accounts with tweets about Merlin via @mentions and Direct Messages. Based on the screening survey responses, we selectively enrolled participants to maximize the diversity of the study sample's domain and AI background (see Tab. 2). We grouped participants based on their survey responses and interview answers.

- *Low-domain*: From "don't know anything about birds" (P11, P12) to "recently started birding" (P7, P8, P13, P14, P16). Participants who selected the latter option typically have been birding for a few months or more than a year but in an on-and-off way, and were able to identify some local birds.
- *Medium-domain*: Have been birding for a few years and/or can identify most local birds (P1, P2, P4, P6, P10, P20).
- *High-domain*: Have been birding for more than a few years and/or do bird-related work (e.g., ornithologist) (P3, P5, P9, P15, P17, P18, P19).
- *Low-AI*: From "don't know anything about AI" (P16, P17) to "have heard about a few AI concepts or applications" (P2, P5, P7, P12, P20). Participants in this group either did not know that the app uses AI (P12, P16) or knew but weren't familiar with the technical aspects of AI (P2, P5, P7, P17, P20).
- *Medium-AI*: From "know the basics of AI and can hold a short conversation about it" (P1, P3, P8, P9, P14) to "have taken a course in AI or have experience working with an AI system" (P4, P10, P15). Participants in this group had a general idea of how the app's AI might work, e.g., it is neural network based and has learned to identify birds based on large amounts of labeled examples.
- *High-AI*: Use, study, or work with AI in day-to-day life (P6, P11, P13, P18, P19). Participants in this group were extremely familiar with AI in general and had detailed ideas of how the app's AI might work at the level of specific data processing techniques, model architectures, and training algorithms.

Note that our referral here and elsewhere to "high-AI background" participants describes their expertise with AI in general, not necessarily with the app's AI. All participants were active end-users of Merlin who could provide vivid anecdotes of when the app worked well and when it did not. Regarding frequency of use, 11 participants used it several times a week, 8 used it once a week, and one used it once a month.

B INTERVIEW PROTOCOL

We began each interview by introducing the study, communicating that we were not affiliated with Merlin's AI development team, and receiving consent for participation. We then asked the participant about their domain and AI background, as well as goals and stakes in their app use. Next, we inquired about the participant's perception of, experience with, and trust in the app. Regarding trust, we adopted Benk et al.'s [3] trust enablement paradigm and asked participants to describe their trust relationships with the app in their own terms. We asked about general attitudes and

perceptions, such as how accurate and trustworthy they find the app, as well as specific instances, such as how they assess the correctness of the app's outputs, and in what circumstances they decide to use the app and not. Scoping down our unit of analysis, from the system as a whole to the "instance" of use, provided a way to gather dynamic data from our participants about which contextual factors they considered during their trust-related decision-making. The goal of this scoping work was to elicit specific contextual factors which may influence the acceptance or rejection decisions of end-users. Finally, we asked whether they would adopt the app in hypothetical high-stakes scenarios with health-related and financial outcomes:

(1) *Sick bird scenario*: Suppose you find a sick bird and take it to the vet. The vet is not sure what bird it is. Would you recommend Merlin to identify the bird species so that the vet can determine the course of treatment?
(2) *Game show scenario*: Suppose you enter a game show where you can win or lose money based on how well you can identify birds from photos or audio recordings. You can only use one resource among Merlin, books (e.g., field guides), the Internet (e.g., search engine, online birder community), and so on. Which resource would you use? Does your answer change depending on certain factors?

We designed these scenarios to introduce high stakes into the AI adoption decision. These scenario-based inquiries allowed us to observe how participants' trust attitudes and trust-related behaviors differ across usage contexts.

C DEFINITION OF TRUST IN OUR STUDY CONTEXT

In this section, we examine if "trust" is the right term for describing participants' attitudes toward the app. Recall from Sec. 2 that trust is defined as an *attitude* and requires *positive expectations* and *vulnerability* in the trustor-trustee relationship [19, 21, 29]. It is easy to see that participants had *positive expectations*: they were actively using the app because they expected it to help them achieve their goal of accurately identifying birds. However, is there *vulnerability* in use of this everyday app for bird identification? We answer yes because the app is used in situations involving *uncertainty of outcomes* and *potentially negative consequences*, satisfying [29]'s definition of *vulnerability*.

First, bird identification is a challenging task that requires the selection of a species among approximately 10,000 existing bird species, some of which are markedly similar to each other. Even though the app has been developed by bird and AI experts and trained on a large database of expert-annotated bird photos and audio recordings, it is not foolproof. There is always *uncertainty* about whether it would return an accurate identification, which participants were aware of. See Sec. 4.1 for detailed accounts of how participants perceived the app's ability and trustworthiness.

There are also *potential negative consequences* when the app makes a misidentification. We heard the following responses when we asked participants what they gain and lose when the app succeeds and fails on the task. As gains, participants mentioned satisfaction of curiosity (All), joy (P1, P7, P9, P12), bird knowledge (P4, P5, P8, P9, P10) and improved birding experience (P1, P2, P3, P4, P10). As losses, although several participants said "*nothing material*" (P1, P3, P11, P12, P13, P15, P16, P17), many expressed that they feel "*disappointed*", even "*frustrated*", when the app fails because they really care about correctly identifying birds and would like to gain accurate results and knowledge from their birding experience (P1, P4, P5, P6, P9, P10, P13, P15, P18, P19, P20). Some noted that misidentifications can lead to people gaining wrong knowledge, (unintendedly) sharing misinformation by reporting wrong bird sightings, and negatively impacting science and conservation efforts (P4, P5).

In summary, there were *positive expectations* and *vulnerability* in participants' use of the app, although there were individual differences in the amount of stakes participants placed in their use. Hence, we conclude "trust" is the right term for describing participants' *attitudes* toward the app.