Int. J. Human–Computer Studies 172 (2023) 102977

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Mitigating knowledge imbalance in AI-advised decision-making through collaborative user involvement

Catalina Gomez [*], Mathias Unberath, Chien-Ming Huang

Department of Computer Science, Johns Hopkins University, MD, USA

ARTICLE INFO

Keywords:
Decision-making
Human–AI interaction
Knowledge imbalance
Trust calibration

ABSTRACT

Integrating artificial intelligence (AI) systems into decision-making tasks attempts to assist people by augmenting or complementing their abilities and ultimately improve task performance. However, when considering recommendations from modern "black box" intelligent systems, users are confronted with the decision of accepting or overriding AI's recommendations. These decisions are even more challenging to make when there exists a significant knowledge imbalance between the users and the AI system—namely, when people lack necessary task knowledge and are therefore unable to accurately complete the task on their own. In this work, we aim to understand people's behavior in AI-assisted decision-making tasks when faced with the challenge of knowledge imbalance and explore whether involving users in an AI's prediction generation process makes them more willing to follow the AI's recommendations and enhances their perception of collaboration. Our empirical study reveals that the involvement of users in generating AI recommendations during a task with notable knowledge imbalance causes them to be more willing to agree with the AI's suggestions and to perceive the AI agent and their collaboration as a team more positively.

1. Introduction

AI-powered systems have the potential to assist humans in real-world tasks, including high-stakes assignments in which complete automation may not be desirable; in such situations, the AI system provides a recommendation and the human user is responsible for making the final decision, commonly known as *AI-advised human decision-making* (Bansal et al., 2019a). A growing number of real-world applications are exploring the use of AI assistance in support of human decision-making, including decision support tools in clinical environments for patient triage (Berlyand et al., 2018), recidivism prediction in criminal justice (Lima et al., 2021), credit assessment, and recommender systems for online retailers, streaming services, and social media (Kunkel et al., 2019; Ngo et al., 2020). While high performance is a required property of assistive AI systems, robustness and explainability are fundamental factors for trustworthy AI (Holzinger, 2021; Holzinger et al., 2022) to avoid undesired outcomes, such as unreliable or unfair decisions. As humans receive assistance from AI systems, it is important to understand how they trust these systems and their recommendations. The success of human–AI teams also depends on how humans interpret and incorporate AI recommendations into tasks to achieve superior joint performance, user satisfaction, understanding of the system, and positive perceptions of the AI system (Smith-Renner et al., 2020).

Multiple factors shape the success of human–AI collaboration. For instance, the timing of AI recommendations has been shown to affect joint decision-making; in particular, the immediate display of an AI recommendation before allowing people to make a decision first can result in anchoring bias (Buçinca et al., 2021). Moreover, users' first impressions based on the correctness of an AI's predictions affect their trust in the system over time (Nourani et al., 2020). Different levels of users' domain or task expertise also affect how people interact with AI systems (Wang and Yin, 2021; Nourani et al., 2020; Gaube et al., 2021; Buçinca et al., 2021; Lai et al., 2021); for example, novice users are more likely to blindly follow AI recommendations, a tendency known as *automation bias* (Nourani et al., 2020). Though ideally expert users are involved in critical decision-making, access to highly specialized individuals may be limited in practice—such as in remote setups or situations with constrained resources available for training experts. Furthermore, the knowledge gap between users and an AI system may vary significantly from task to task. In this work, we use the term *knowledge imbalance* to refer to situations in which an AI system has a substantially higher solo accuracy than the user in completing the task at hand; for instance, a large knowledge imbalance can be observed when an AI system provides assistance in a "superhuman" task—i.e., a task that is difficult for humans because there are no clear directions on how to solve or complete it. As an example in the clinical domain, AI

[*] Corresponding author.
E-mail address: cgomezc1@jhu.edu (C. Gomez).

https://doi.org/10.1016/j.ijhcs.2022.102977
Received 5 August 2022; Received in revised form 29 November 2022; Accepted 5 December 2022
Available online 13 December 2022

systems may leverage visual information from histopathology slides to generate a diagnosis that usually requires a molecular test in addition to the manual inspection of specialists (Diao et al., 2021; Chen et al., 2020).

Various challenges are associated with AI-advised human decision-making in a knowledge imbalance setup. In particular, appropriate trust calibration is difficult to achieve, as laypeople may not be able to identify when the AI makes a correct or incorrect recommendation (Nourani et al., 2020), thus prompting the potential adoption of simple and less collaborative heuristics (Bansal et al., 2021). Even high quality machine learning models are vulnerable to poor decisions under unexpected data variations, highlighting the importance of users' assessments of model outcomes (Holzinger, 2021). The most common AI-advised human decision-making paradigm involves an AI presenting a recommendation and its user deciding whether or not to follow it (Bansal et al., 2019a); such one-way interaction does not allow the user to provide any input to the model, whether to refine its output, provide additional labels, or improve it in another way. In this one-way interaction, if the AI system is sufficiently accurate, users can achieve superior task performance if they accept all of its recommendations; however, they might lose agency in the decision-making process and not perceive their interactions with the AI positively. Alternatively, users could reject all of the AI's recommendations and find themselves facing undesirable task outcomes.

As motivated by the human-in-the-loop approach to model-building processes in interactive machine learning (Amershi et al., 2014; Wall et al., 2019), in this work we explore the role of interactivity in AI-advised human decision-making when there exists a notable knowledge imbalance between the human and the AI agent; specifically, we let users provide information that they can understand from the task as features to guide the AI model's prediction. Our central hypothesis is that involving users as active contributors in the process of generating AI recommendations will mitigate the aforementioned challenges associated with knowledge imbalance in AI-advised human decision-making. Next, we provide a brief summary of the relevant prior research that motivated this work.

2. Related work

2.1. AI-assisted decision-making

AI assistance can support human users in different types of tasks, including (1) familiar tasks, such as text completion (2) tasks that can be learned with practice, such as predicting an outcome (e.g., student admissions decisions), (3) tasks that require specialized knowledge, such as fine-grained image classification and interpretation in medicine. As people make the final decisions and determine whether or not they follow the AI's recommendations, appropriate trust calibration requires users to catch errors either by using their own knowledge and experience or via the presentation of information or cues by the AI that allow them to create an accurate mental model (Bansal et al., 2019a); however, such information might not be available or directly accessible in practice (e.g., model performance in a new data sample). Presenting the model's recommendations along with explanations has received considerable attention in an attempt to make decision making processes more transparent and promote trust (Holzinger et al., 2019). The lack of understanding how these models work, how the provided information is processed, and whether they rely on the right information might hinder people's ability to identify when a model is being unfair (Angerschmid et al., 2022). Certain explanation types and visualization techniques to present additional model information have been identified as more effective to enhance people's perception of fairness (Dodge et al., 2019; Van Berkel et al., 2021). However, previous studies have demonstrated that current approaches to explaining machine learning predictions can be easily misunderstood by laypeople, affecting their performance and trust (Lai and Tan, 2019) or simply having no observed benefit (Nourani et al., 2021; Cheng et al., 2019).

Multiple factors can influence users to follow incorrect recommendations from an AI agent. When users have high confidence in the task but no information about the model's performance is presented, the agreement between the users' own decisions and the AI model's predictions seems to affect the positive perception of the model's performance, motivating users to rely more on the model while ignoring the correctness of its predictions (Lu and Yin, 2021). Furthermore, even when users perform well in a task and have the ability—or can develop the intuition—to identify when the AI's recommendations may be off, they follow incorrect recommendations anyway (Suresh et al., 2020). A similar behavior has been observed when users must make decisions under covariate shift, wherein the model's performance is likely to decline; despite being aware of changes in the data and their self-performance, users still tend to rely on the model (Chiang and Yin, 2021). Meanwhile, studies with more experienced users have shown that experts can determine when to ignore erroneous recommendations, even when overriding them requires more effort (De-Arteaga et al., 2020; Levy et al., 2021; Gaube et al., 2021). Even though humans make the final decision when they receive assistance from an AI agent, their involvement in the decision-making process should go beyond determining whether or not to follow a recommendation and processing the provided supporting evidence (if available) in order to decrease negative trust outcomes.

2.2. User involvement in AI-assisted decision-making

Besides accepting or rejecting AI recommendations, a user's role in a decision-making task can be expanded to encourage specific behaviors or gain further benefits. For instance, when interacting with imperfect algorithms in a forecasting task, people prefer to use an AI model more often and report higher satisfaction if they can modify its forecasts, regardless of the amount by which they are able to adjust the outcomes (Dietvorst et al., 2018). Cognitive forcing functions have been studied to improve user engagement in analytical thinking when considering AI recommendations and explanations, thereby reducing overreliance (Buçinca et al., 2021); by forcing users to make an initial decision or to wait until the AI's suggestion is presented, it is implied that users should have some domain knowledge to engage with the task before they have access to the AI's suggestion. Interactive machine learning further allows people to modify or guide the model's outcomes, which can ultimately improve performance. Users can react to model predictions with instance-level (correcting or confirming predictions) or feature-level feedback (denoting features indicative of each outcome class). A controlled experiment that used a classification task with support for user feedback demonstrated positive effects on user satisfaction (frustration, trust, and model acceptance) and increased expectations that models would improve (Smith-Renner et al., 2020).

However, even though users had the chance to give feedback to the model by providing task-relevant inputs, this interaction occurred after a prediction had already been made, which may have biased users' responses. As another example, medical image retrieval systems can be enhanced with interactive mechanisms that support manual refinement, as these would allow more agency for specialists to guide outcome generation and acquire more in-depth insights about the current model (Cai et al., 2019). In this work, we study user involvement in aiding the generation of AI recommendations by providing task-relevant input and how such involvement may affect human–AI collaboration under a knowledge imbalance scenario.

2.3. User characterization in AI-assisted decision-making

The human-centered approach to designing human–AI interactions has motivated the comprehensive understanding of end users' abilities to achieve successful human–AI collaboration. Overall, users' expertise has been categorized based on their familiarity with AI technologies, as it may be challenging for laypeople to understand algorithmic decision-making systems (Cheng et al., 2019) and create appropriate mental models of said systems (Bansal et al., 2019a). Recently, the consideration of end users has received particular attention in the design of explainable AI systems (Chen et al., 2022; Eiband et al., 2018; Schoonderwoerd et al., 2021), as most of the current techniques are based on the intuition of researchers (Miller, 2019). To alleviate the disconnect between AI developers and the stakeholders targeted in the design process of such systems, one multidisciplinary framework has considered different design goals and appropriate evaluation metrics for each user group: AI novices, data experts, and AI experts (Mohseni et al., 2021); similarly, a granular characterization of stakeholders has been proposed in terms of the knowledge they may possess (formal, instrumental, and personal) and the context in which it manifests (Suresh et al., 2021). In addition to expertise considerations, user needs have been distilled based on long-term goals, short-term objectives, and the tasks required to accomplish them to create a framework for better understanding end users and identifying opportunities in interpretability literature (Suresh et al., 2021). Another guidance tool to inform design practices for explainable AI represented user needs as prototypical questions that they may ask about explainable systems (Liao et al., 2020). In our work, we consider end users' task expertise as another potential dimension to characterize them and focus particularly on how to achieve successful AI interactions for a group of users with no preexisting domain knowledge.

2.4. Assistance under knowledge imbalance

One of the goals of augmenting human decisions with AI assistance is to improve the joint performance of human–AI collaboration (Bansal et al., 2019b). When people have a significantly poorer performance than an AI model—i.e., when users lack expertise or the task at hand is considered "superhuman"—automation bias can result in users blindly following the AI's recommendations while still achieving high accuracy, provided the model is sufficiently accurate. Thus, performance improvement should not be the only factor to consider for appropriate trust calibration (Wang and Yin, 2021; Bansal et al., 2021), especially when there is a knowledge imbalance between the advisor (e.g., the AI) and the advisee (e.g., a non-expert user). This knowledge gap has been considered when evaluating AI-assisted decision-making with different participant populations, including clinicians who are not radiologists reading X-ray images (Gaube et al., 2021); non-algorithmic experts performing predictive tasks (Cheng et al., 2019); novice users in exploratory tasks (Nourani et al., 2021); and non-experts in the entomology (Nourani et al., 2020), botany (Yang et al., 2020), and nutrition domains (Buçinca et al., 2021). Explaining the model's output has been proposed to help users understand the reasoning behind a particular recommendation; however, the effects of such explanations in a task for which users had little domain expertise in the first place were inconclusive in a previous study (Wang and Yin, 2021). Rather than "fixing" the problem of users' expertise or lack thereof, another study analyzed the different degrees of complementary expertise between an AI agent and its human user, from a complete overlap in expertise to a case where they perfectly complemented each other in their skills to identify categories in a classification task (Zhang et al., 2022). Even though humans can easily identify errors made by an AI in tasks in which they are already experts, better trust and reliance measures in cases of perfect complementarity suggest that users ignore such errors in their trust calibration process in these situations. Therefore, more research should focus on situations in which the user has limited expertise and requires assistance from the AI to complete a task, i.e., a knowledge imbalance scenario.

Given that non-experts cannot identify potential errors and that providing additional information can be misleading or at best ambiguous, new strategies should be explored to account for different levels of expertise in target populations. Although different levels of domain expertise have been previously considered in AI-assisted decision-making, users could still somehow judge the presented AI recommendations because they had some preexisting familiarity level with the task, they could gain experience through practice, or they felt they knew something about the task already. As it remains unclear how users trust and perceive an AI system in tasks with a high knowledge imbalance, we endeavored to explore this specific situation.

3. Methods

In this section, we describe the user study we conducted to understand how people who lack the necessary knowledge of an experimental task may work with an AI system to complete that task.

3.1. Hypotheses

We investigated how different paradigms of human–AI joint decision-making under a knowledge imbalance condition affect trust calibration, behavioral indicators of trust, and user perception of the AI agent, with the following hypotheses:

- H1: Participants will increasingly follow AI suggestions when they are more involved in their interaction with the agent than when their interaction is limited to only directly asking for a suggestion. This hypothesis is informed by previous research on strategies to reduce algorithmic aversion, which indicated that people's responses are closer to the model's forecasts when they can modify the model's output (Dietvorst et al., 2018). In our study, users had an active role in modifying the attributes that guided the AI's prediction generation process in the interaction paradigms with higher levels of user involvement.
- H2: Participants' abilities to distinguish when to trust or distrust the AI agent will improve when they are more involved in the AI interaction (i.e., they are less susceptible to overtrust and undertrust). AI explanations have been hypothesized to foster trust calibration (Zhang et al., 2020; Ribeiro et al., 2016) because they allow people to understand the reasoning behind a prediction. In our study, users were involved in the AI's prediction generation process by providing relevant features, adding transparency to the AI's internal processing.
- H3: The development of trust in the AI agent will be reflected in users' behavior during their interactions with it. The observation of behavior when interacting with intelligent systems has been used as an alternative to measure trust (Papenmeier et al., 2022). Related to H1, behavioral indicators of trust will increase when users are more involved in their interactions with the agent.
- H4: Users' perceptions of their trust in, their collaboration with, and the capabilities of the AI agent will improve when they are more involved in their interactions. Empirical evidence from previous work on the perception of interactive AI systems suggests that (1) greater user involvement during a labeling process improves users' comfort levels and increases the perceived capabilities of the system (Mahmood et al., 2021) and (2) feature-level feedback in an interactive setup improves trust in and acceptance of machine learning systems (Smith-Renner et al., 2020).

3.2. Experimental task and study design

To test our hypotheses, we designed a user study in which participants were asked to complete an image-based bird classification task with the assistance of an AI agent. Images were selected from the Caltech-UCSD Birds 200 dataset (Welinder et al., 2010); we carefully selected images that corresponded to uncommon birds and so that the names of the birds depicted were not associated with their physical appearance (e.g., blue-headed vireo, red-winged blackbird) to avoid

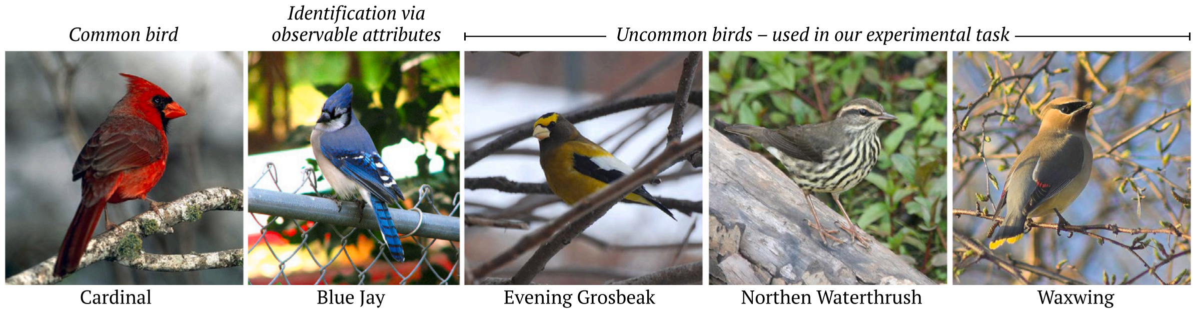

Common bird
Identification via observable attributes
Uncommon birds – used in our experimental task
Cardinal
Blue Jay
Evening Grosbeak
Northen Waterthrush
Waxwing

Fig. 1. Example selected bird images and their corresponding categories. We classified "common" birds as species more likely to be recognized by laypeople. The next category was comprised of species (i.e., blue jays) that could be validated via their observable attributes. The three rightmost images serve as examples of less common birds used in our study that were determined to be harder to classify.

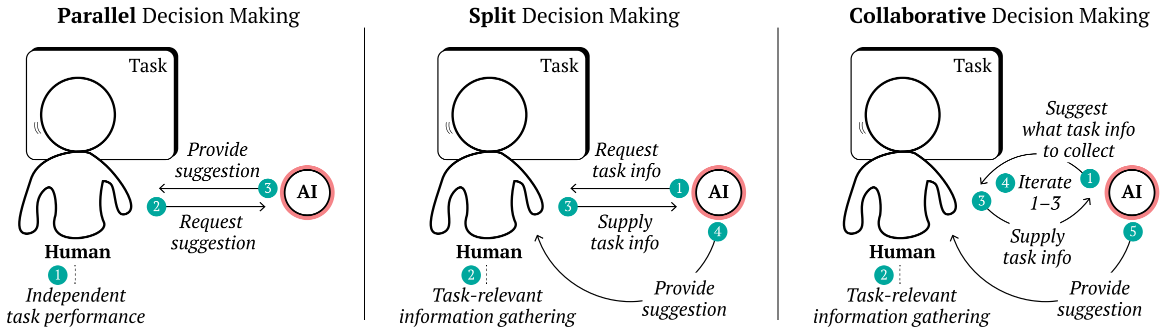

Parallel Decision Making
Task
Provide suggestion
3
AI
2
Request suggestion
Human
1
Independent task performance
Split Decision Making
Task
Request task info
1
AI
3
Supply task info
4
Human
2
Task-relevant information gathering
Provide suggestion
Collaborative Decision Making
Task
Suggest what task info to collect
1
4
Iterate 1–3
3
AI
Supply task info
5
Human
2
Task-relevant information gathering
Provide suggestion

Fig. 2. A schematic representation of our interaction paradigms with different levels of user involvement. The numbers correspond to the steps that take place while interacting with the AI agent.

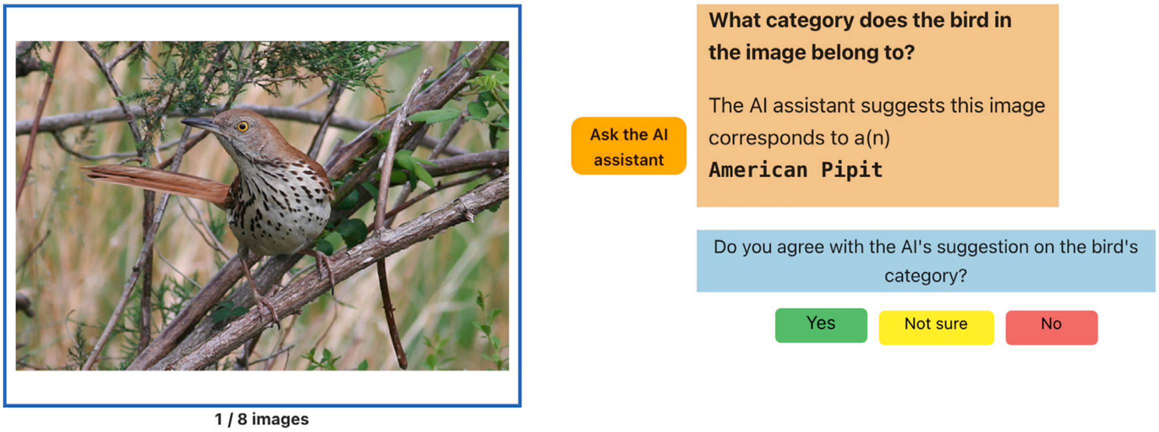
giving participants potential cues by which to judge the AI's suggestions (see examples in Fig. 1). Additionally, the images chosen contained only one bird each, although they were diverse in terms of lighting, background, pose, and scale.

We designed three *interaction paradigms* with different levels of user involvement (as illustrated in Fig. 2) to explore AI-assisted decision-making in a task with a high knowledge imbalance:

- **Parallel decision-making (PDM)**: This paradigm represents conventional human–AI interactions (Bansal et al., 2019a,b) wherein the user and the AI make decisions in parallel; the AI only offers its recommendations upon request. In our implementation, the user requested an AI recommendation after having the opportunity to work on the task alone (e.g., viewing the task image). This paradigm is commonly implemented in applications such as juridical assistance regarding recidivism, risk assessment, and assisted medical diagnosis.
- **Split decision-making (SDM)**: In this paradigm, the user and the AI split the load of decision-making; for example, the user may prepare task-relevant information according to their understanding of the task in order for the AI to make a recommendation. In our implementation, the user was presented with a task image and asked to describe the physical features of the bird (e.g., colors and patterns of body parts) that they considered relevant in order for the AI model to make an accurate prediction. If needed, the AI could ask for additional input. Once the user described at least three physical attributes, the AI would suggest the bird category corresponding to the provided attributes.
- **Collaborative decision-making (CDM)**: In this paradigm, the user and the AI collaborate more closely throughout the decision-making process by iterating over the task-relevant information necessary for the AI to provide a final prediction. In our implementation, the AI suggested which attributes were most relevant for the user to describe and made a final bird category recommendation based on the user's input. If needed, the AI could ask for additional input. Once the user entered at least three attributes, the AI would present its suggestion of the bird category to which the featured bird belonged. This paradigm contrasts with SDM, where the user and AI take turns in contributing to the decision-making task.

To complete the bird classification task in the PDM interaction paradigm, participants merely requested the AI's recommendation after viewing the task image by clicking the "Ask the AI assistant" button, as shown in Fig. 3. For the SDM and CDM interaction paradigms, participants first had to describe the bird in the presented image by selecting relevant physical attributes (from eighteen body parts and color, pattern, and shape characteristics) and their corresponding values from a fixed set of options provided via a drop-down menu; they had access to the detailed bird part information while describing the image. Once they selected a body part, characteristic, and value, they could add attributes one at a time to the current cumulative description. Fig. 4 illustrates the interactive fields in our web interface that allowed users to provide a bird description in the SDM interaction paradigm; the same format was used for the CDM condition. The submitted body-part–characteristic pair was removed from the drop-down menu options following its selection to avoid duplicate descriptions and to reduce the number of attributes left to describe. Users had the option to delete the last added attribute or to restart their entire description at any time. Every time an attribute was added or removed, a GIF mimicking the AI's processing was displayed for 2 s.

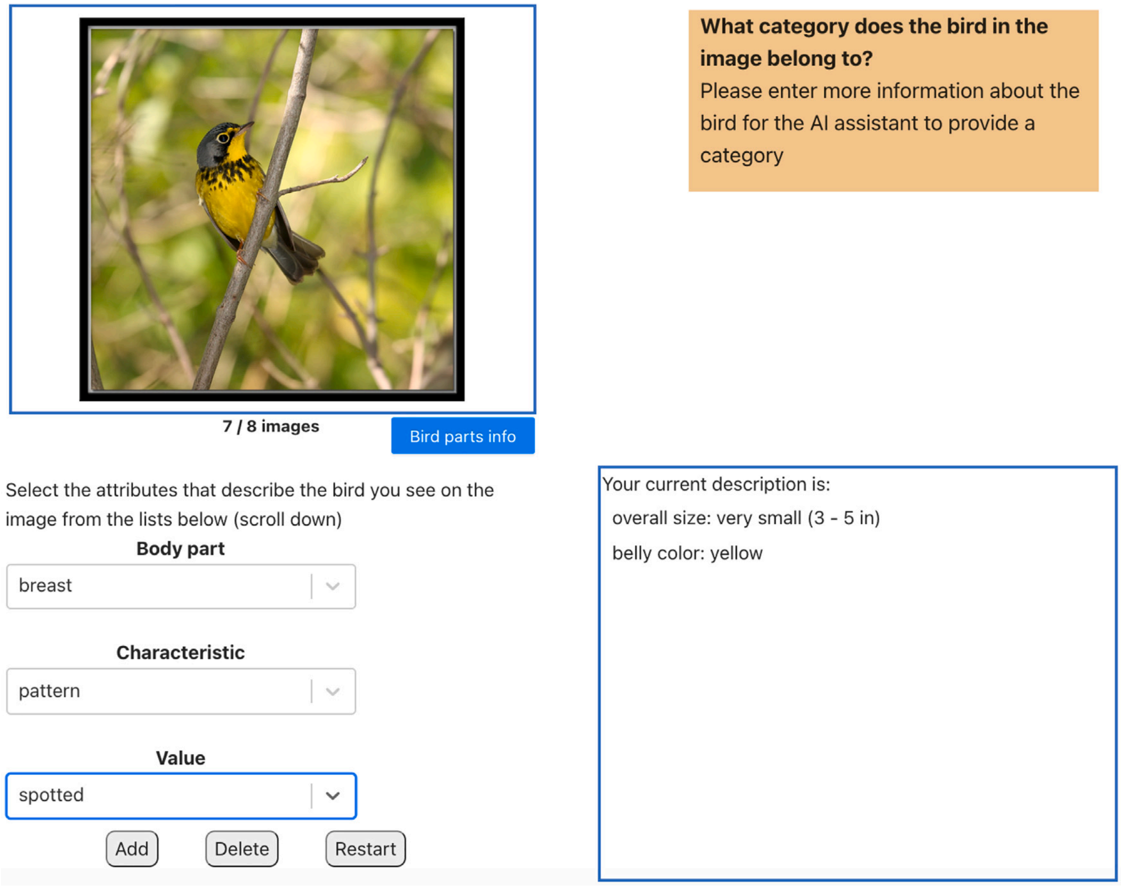
The AI agent could request the user to describe more of the bird's physical attributes in order to provide a final category (see the top right box in Fig. 4). We conducted initial tests of the user interface and defined the minimum description length for the AI to provide a bird category as three attributes and fixed it as such for all the images assigned to the SDM and CDM interaction paradigms. Thus, even if users added more items to the description or modified the minimum description required for the model to display its recommended category, they would still get the same AI suggestion because we simulated


What category does the bird in the image belong to?

The AI assistant suggests this image corresponds to a(n) American Pipit

Do you agree with the AI's suggestion on the bird's category?

Fig. 3. Example of our web interface for the PDM interaction paradigm. Users could request the AI's bird category suggestion by clicking the “Ask the AI assistant” button.


What category does the bird in the image belong to?
Please enter more information about the bird for the AI assistant to provide a category

Select the attributes that describe the bird you see on the image from the lists below (scroll down)

Body part: breast

Characteristic: pattern

Value: spotted

Your current description is:
overall size: very small (3 - 5 in)
belly color: yellow

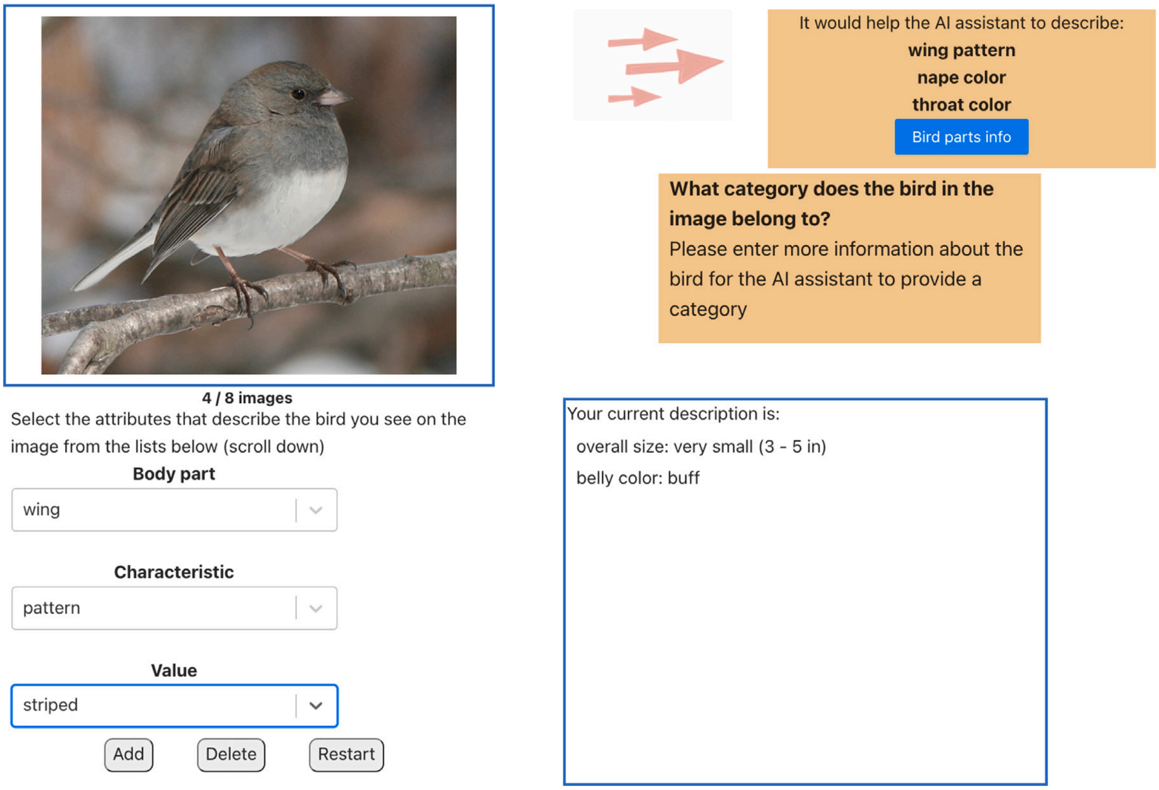
Fig. 4. Example of our web interface for the SDM interaction paradigm. Suggestions on attributes to describe were displayed next to the bird image and pointed out with large arrows. Users provided physical attributes using the drop-down menu below the image and the current description was updated in the box to the right.

the outcomes of the model and it was not actually considering the users' inputs. This design choice isolates perceptions of the interaction paradigms from how informative are the user's bird descriptions for the model to generate a correct prediction, and in this way kept the same model's performance in all the three interaction paradigms.

In the CDM interaction paradigm, the AI's suggested attributes to include in the final bird description were presented as body-part–characteristic pairs, as shown in the top right box in Fig. 5. The AI agent presented three suggestions at a time. Each time the user described one of the suggested body-part–characteristic pairs, the AI agent updated its suggestions with a pair from a list of predefined salient attributes that were manually selected for each bird image; new suggestions did not include attributes that had been already described by the user so as to emulate the AI's consideration of the current cumulative description in its processing. We intentionally delayed the presentation of a new suggestion by one second to imitate the AI processing the new information.

In all three interaction paradigms, the presentation of the AI's recommended bird category was followed by a question asking whether or not the user agreed with that recommendation; participants chose among three responses: *Yes, Not sure,* or *No,* as shown in Fig. 3.

AI suggestion generation. To mimic a real classification model, the AI's bird category suggestions were predetermined and had an overall classification accuracy of 63% (15 out of 24 images), which was kept



It would help the AI assistant to describe:
wing pattern
nape color
throat color

What category does the bird in the image belong to?
Please enter more information about the bird for the AI assistant to provide a category

4 / 8 images
Select the attributes that describe the bird you see on the image from the lists below (scroll down)

Body part: wing

Characteristic: pattern

Value: striped

Your current description is:
overall size: very small (3 - 5 in)
belly color: buff

Fig. 5. Example of our web interface for the CDM interaction paradigm. Suggestions on attributes to describe were displayed next to the bird image and pointed out with large arrows. Users provided physical attributes using the drop-down menu below the image and the current description was updated in the box to the right.

the same for all interaction paradigms (i.e., five correct and three incorrect predictions for eight images). By asking participants to evaluate accurate and inaccurate recommendations, we were allowed to study appropriate trust outcomes under knowledge imbalance. We manually defined the incorrect bird category suggestions by considering birds with appearances similar to those in the correct category, ensuring that the new categories did not contain any embedded physical attributes in the bird name.

Each participant was exposed to all three interaction paradigms as in a within-subjects design. The order of the interaction paradigms and the images assigned within each condition were randomized. We administered our user study using a custom web application developed with React JS, Express.js, and Firebase.

We conducted an initial online pilot study and, based on the resulting data quality, decided to conduct our final study in person. Our main concerns with the data collected from the online study were participants always choosing the same response or answering randomly and that they may have relied on additional help (i.e., Google search) to validate the AI's recommendations; thus, we conducted our final study in person.

3.3. Measures

We used a set of objective and subjective metrics to understand how people interact with an AI agent when a high knowledge imbalance exists between the collaborating parties. Our objective metrics aimed to capture participants' levels of agreement with the AI agent, whether their trust in the agent's recommendations was properly calibrated, and the presence of behaviors associated with trust in the agent. Our subjective metrics aimed to capture participants' perceptions of the AI agent while they completed the task.

3.3.1. Objective metrics

- **Agreement with the AI.** To evaluate users' willingness to follow the AI's recommendations, we used participants' final answers regarding their concurrence with the bird category suggestions presented by the AI agent; the answer options were: *Yes, No,* and *Not sure.*
- **Overtrust and undertrust.** We evaluated users' trust calibration outcomes when accepting or rejecting recommendations from the AI agent via two dependent variables, *overtrust* and *undertrust,* as informed by the trust outcome variables used in previous work (Wang and Yin, 2021). In our analysis, we considered that a decision resulted in overtrust when users answered *Yes* to incorrect recommendations and undertrust when users answered *No* or *Not sure* to correct recommendations. Each variable (overtrust or undertrust) was binary; i.e., users either did or did not over/undertrust the AI's recommendation for each image.
- **Behavioral indicators of trust.** We defined two metrics that take into account the bird attributes provided by participants in the SDM and CDM interaction paradigms. The first metric compares the number of bird attributes included in the final description for each image in the SDM and CDM conditions. Since the minimum number of attributes for the AI to provide a bird category prediction was three for both paradigms, any additional attributes provided by the participants could be viewed as a lack of confidence in the current bird category prediction; therefore, we associated descriptions with fewer total items with more trust in the AI agent. The second metric is the percentage of physical attributes in the final description provided by the user that were originally suggested by the AI in the CDM paradigm; a higher percentage indicates that participants preferred to use the AI's attribute suggestions rather than proposing their own, which could be associated with higher levels of trust in the AI agent to complete the task.

- **Decision time.** This metric represents the time users spent considering the AI's bird category recommendations. Time allocation strategies have been explored in AI-assisted decision-making tasks (Rastogi et al., 2022) to reduce potential biases when evaluating AI recommendations, with longer decision times helping the decision-maker sufficiently adjust away from a biased anchor. We measured decision time as the time from the appearance of the AI's bird category prediction to the moment the participant chose their final response agreeing or disagreeing with it. Decision time can be an indirect measure of trust—the more hesitant the participant is, the more time it takes them to make a decision on whether or not they want to follow the AI's recommendation; therefore we associated shorter decision times with more trust in the AI agent.

3.3.2. Subjective metrics

We assessed users' perceptions of the AI agent through eight subjective statements to be rated using a 5-point Likert scale, with 1 being "Strongly disagree" and 5 being "Strongly agree". The statements probed participants' perceived trust in and reliance upon the AI, the perceived usefulness of those recommendations, the perceived knowledge of the AI agent regarding the task, whether they perceived the AI's suggestions as questionable, their perceived sense of teamwork, the AI agent's perceived contribution toward completing the task, and its perceived contributions with respect to the participants'. We additionally asked participants to subjectively rate their ability to complete the task successfully without the AI agent. The complete set of statements is listed in the Appendix.

3.4. Experimental procedure

Upon agreeing to participate in the study with informed consent, participants were asked via our user interface for their demographic information, including their age range, gender, education level, and familiarity with AI technology. They then took a pre-study test to assess their knowledge of birds; in this test, participants had to provide a category for five birds, three common and two uncommon. The instructions for the main task were then presented to them and they reviewed useful information about birds' body parts. Besides, users were required to complete a practice bird description task to familiarize themselves with the information available in the drop-down menus before moving on.

In the main task, participants classified 24 images within 24 different bird categories, with one image representing each category. The order in which the three interaction paradigms were presented to each participant was random, and the distribution of images (eight) per interaction paradigm was random as well. For each trial, the user was presented with a bird image and additional instructions (depending on the interaction paradigm) to prompt the AI's bird category recommendation. Once the AI's suggestion was presented, we asked users if they agreed with that category choice; they selected among three possible responses: *Yes, Not sure,* or *No*. Participants were allowed to change their response before moving to the next image and we recorded their response history. This process was repeated for all eight images in each interaction paradigm. After completing all the trials within one interaction paradigm, participants were asked to respond to a questionnaire about their experience with the AI agent in that paradigm. In the final questionnaire, after the users had been exposed to all three interaction paradigms, they were asked to choose which paradigm they would prefer for future use.

This user study was approved by our institutional review board. On average, participants took 40 min to complete the study and were compensated with a $10 gift card.

3.5. Participants

We recruited 28 participants (13 female, 13 male, and 2 self-identified as otherwise gendered) through convenience sampling from a local university community. The age range distribution was $n = 14$ between 18–24 years, $n = 10$ between 25–29 years, $n = 3$ between 30–34 years, and $n = 1$ for 50 years or older. Most of the participants were current graduate students ($n = 17$), followed by those who had completed graduate school ($n = 5$), those who had graduated from college ($n = 3$), and current undergraduate students ($n = 3$). Participants had an average rating of 4.12 ($SD = 0.88$) with regard to their familiarity with AI technology on a scale from 1 to 5, where 1 denotes "Extremely unfamiliar" and 5 denotes "Extremely familiar".

4. Results

We used mixed effects models for each dichotomous variable to account for participant-level effects and the repeated measure experimental design. Similarly, for the continuous variables, we used one-way repeated measure analysis of variance (ANOVA) models. All post-hoc pairwise comparisons were conducted using pairwise paired t-tests with Bonferroni correction and we report the adjusted p-values. Paired t-tests were used when the comparisons involved two groups. Assumptions of normality and homogeneity of variance were checked with Shapiro–Wilke and Levene's test, respectively, and outliers were identified using the interquartile range, removing extreme values for the statistical tests (Lazar et al., 2017). For all the statistical tests reported below, $p < .05$ is considered as a statistically significant effect. We followed Cohen's guidelines (Cohen, 1988) on effect sizes and considered $\eta_p^2 = 0.01$ a small effect size, $\eta_p^2 = 0.06$ a medium effect size, and $\eta_p^2 = 0.14$ a large effect size. For Cohen's or Hedge's index, 0.2 is considered small, 0.5 medium, and 0.8 a large effect.

4.1. Validation of knowledge imbalance

To verify that there was a large knowledge gap between the AI system and the human participants, we administered a pre-study test on participants' knowledge of bird categories, in which the average number of correctly identified birds was 0.79 ($SD = 0.92$) out of 5. Additionally, we did not provide any training or immediate feedback to users regarding the AI's performance or the correctness of each trial during the main experiment to prevent any learning effects and keep the knowledge imbalance. The knowledge gap was further confirmed by users' high rate of agreement with the statement, "I would not have been able to successfully complete the task without the AI": 4.11 ($SD = 1.13$), 3.96 ($SD = 1.14$), and 4.32 ($SD = 0.98$) on a 5-point Likert scale in the PDM, SDM, and CDM interaction paradigms, respectively.

4.2. Distribution of user responses to AI recommendations

We analyzed the distribution of users' responses regarding their agreement with the AI's bird category recommendations in the 672 total trials (eight trials for three interaction paradigms for each of the 28 participants). Table 1 presents the distribution of responses for each interaction paradigm and the overall count. In both the PDM and CDM interaction paradigms, *Not sure* was the most common response, followed by *Yes*, and then very few *No* responses. In the SDM interaction paradigm, the number of *Yes* and *Not sure* responses was equal and users did not respond *No* in any trial. Overall, there were 296 *Yes,* 5 *No,* and 371 *Not sure* responses in the 672 trials of all participants and in all paradigms. We note two observations from the distribution of responses: (1) participants were less likely to answer *Yes* than *Not sure* in the PDM interaction paradigm and (2) participants barely answered *No* in all three interaction paradigms, presumably because they had little knowledge of bird classification in the first place. As a result, in the following analyses, we combine the *No* and *Not sure* responses to represent disagreement with the AI's recommendations, while *Yes* responses represent agreement; in other words, we considered agreement with the AI's recommendations as a binary dependent variable.

Table 1
Distribution of participant agreement with the bird category suggested by the AI system for all trials in each interaction paradigm.

Interaction paradigm	Yes	No	Not sure	Total
Parallel decision-making	76	3	145	224
Split decision-making	112	0	112	224
Collaborative decision-making	108	2	114	224
Total	296	5	371	672

Table 2
Generalized linear models for agreement with the AI agent and interaction paradigm. SE denotes standard error.

Mixed effects logistic regression for agreement with the AI						
Outcome	Model term	β	SE	z	p	Odds ratio
Agree (answered *Yes*)	Intercept	−0.94	0.31	−3.02	.003	–
	SDM	0.91	0.23	4.00	<.001	2.48
	CDM	0.81	0.23	3.57	<.001	2.25

Table 3
Generalized linear models for overtrust in the AI's recommendations and interaction paradigm. SE denotes standard error.

Mixed effects logistic regression for overtrust outcomes						
Outcome	Model term	β	SE	z	p	Odds ratio
Overtrusted	Intercept	−0.72	0.32	−2.24	.025	–
	SDM	0.67	0.35	1.91	.056	1.95
	CDM	0.73	0.35	2.08	.038	2.07

4.3. Do different levels of user involvement affect non-expert users' agreement with AI recommendations when interacting with an AI agent?

Table 2 exhibits the detailed results of a mixed effects logistic regression (dependent variable [DV]: user responses—agree with the AI or not; independent variable [IV]: interaction paradigm—PDM, SDM, or CDM), where the PDM paradigm was set as the baseline group. The results suggested a significant positive influence of the SDM ($p < .001$) and CDM ($p < .001$) interaction paradigms on users' agreement with the AI agent, suggesting that participants were significantly more likely to agree with the agent in those conditions than in the PDM interaction paradigm. More specifically, it was found that the odds of agreeing with the AI agent increased by 148% and 125% for participants in the SDM and CDM interaction paradigms compared to the standard PDM, respectively.

4.4. Do different levels of user involvement affect non-experts' development of overtrust and undertrust when interacting with an AI agent?

4.4.1. Overtrust

To evaluate overtrust, we only considered the trials in which the AI agent provided incorrect recommendations—i.e., three of eight trials in each interaction paradigm, or 252 trials in total. We measured the effect of the interaction paradigm (IV) on developing overtrust in the AI agent using a mixed effects logistic regression with a binary DV (whether or not participants overtrusted the AI agent when it was incorrect). Table 3 summarizes the detailed outcomes of the regression. We observed a positive influence of the SDM (41/84 trials with overtrust) and CDM (42/84 trials with overtrust) interaction paradigms on overtrusting the AI agent, but only the latter influence was significant ($p = .038$) compared to the PDM paradigm. This finding suggested that participants in the CDM interaction paradigm had 107% more odds of overtrusting the AI agent by accepting its incorrect recommendations than those in the standard PDM paradigm.

Table 4
Generalized linear models for undertrust in the AI's recommendation and interaction paradigm. SE denotes standard error.

Mixed effects logistic regression for undertrust outcomes						
Outcome	Model term	β	SE	z	p	Odds ratio
Undertrusted	Intercept	1.07	0.36	2.94	.003	–
	SDM	−1.05	0.29	−3.57	<.001	0.35
	CDM	−0.84	0.29	−2.89	.004	0.43

4.4.2. Undertrust

To evaluate undertrust, we only considered the trials in which the AI agent provided correct recommendations—i.e., five of eight trials in each interaction paradigm, or 420 trials in total. We measured the effect of the interaction paradigm (IV) on developing undertrust in the AI agent using a mixed effects logistic regression with a binary DV (whether or not participants undertrusted the AI agent when it was correct). Table 4 summarizes the detailed outcomes of the regression. The coefficients of the SDM (69/140 trials with undertrust, $p < .001$) and CDM (74/140 trials with undertrust, p = .004) interaction paradigms suggested a significant negative influence on undertrusting the AI agent when compared with the PDM paradigm, which means that participants were less likely to undertrust the AI agent in the former conditions. More specifically, participants in the SDM interaction paradigm had odds of undertrusting the AI agent that were 0.35 of the odds of those in the PDM paradigm. Likewise, participants in the CDM interaction paradigm had odds of undertrusting the AI agent that were 0.43 of the odds of those in the PDM paradigm. This observation also aligns with the higher likelihood that participants followed the AI's recommendations in the SDM and CDM interaction paradigms regardless of correctness; as participants were less likely to reject potentially correct AI recommendations, they were therefore less likely to undertrust the AI agent in general.

4.5. How may different interaction paradigms shape user contribution to joint decision-making?

A paired t-test comparing the number of provided attributes between the SDM and CDM interaction paradigms found a significant difference in the average number of attributes described in each condition ($t(219) = 3.35$, $p < .001$, Hedge's $g = 0.26$). On average, users described more attributes in the SDM interaction paradigm ($M = 3.66, SD = 1.05$) than they did in the CDM interaction paradigm ($M = 3.41, SD = 0.78$), as shown in Fig. 6. Furthermore, in the CDM interaction paradigm, participants followed on average 2.70 ($SD = 0.74$) suggestions from the AI agent regarding which attributes to describe, and these represented on average 84.44% ($SD = 25.54$) of the attributes present in the final bird descriptions.

4.6. Do different levels of user involvement affect non-experts' decision time when considering an AI agent's recommendations?

We used a one-way repeated measure ANOVA where the interaction paradigm was set as a fixed effect and participants as a random effect. Fig. 6 visualizes our results for decision time. The average decision time (in seconds) was longest for the PDM interaction paradigm ($M = 7.50, SD = 3.48$), followed by the SDM ($M = 5.71, SD = 2.69$) and then the CDM ($M = 5.36, SD = 2.09$) paradigms. We found a significant main effect of interaction paradigm on decision time, ($F(2, 54) = 20.66, p < .001, \eta_p^2 = .430$), and further pairwise comparisons using the Bonferroni correction revealed that decision time was significantly higher in the PDM paradigm than in the CDM ($p < .001$) and SDM ($p < .001$).

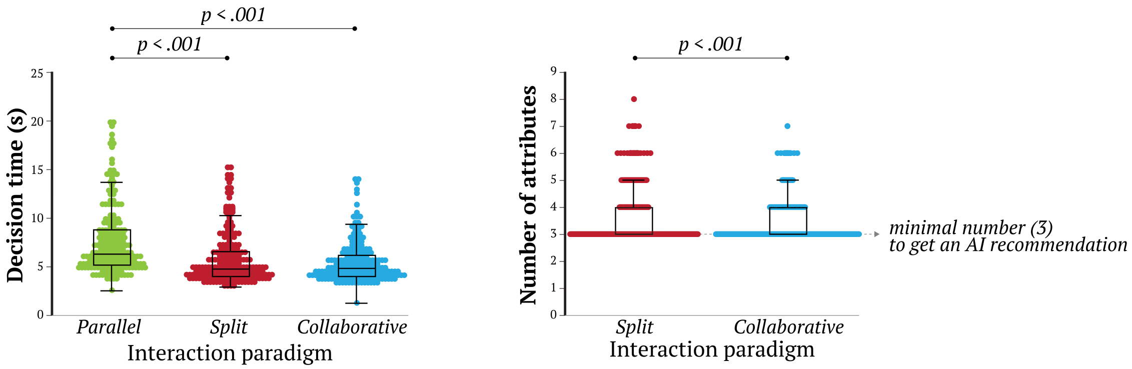


Fig. 6. Box and whisker plots for decision time across all interaction paradigms (left) and description length in the SDM and CDM interaction paradigms (right).

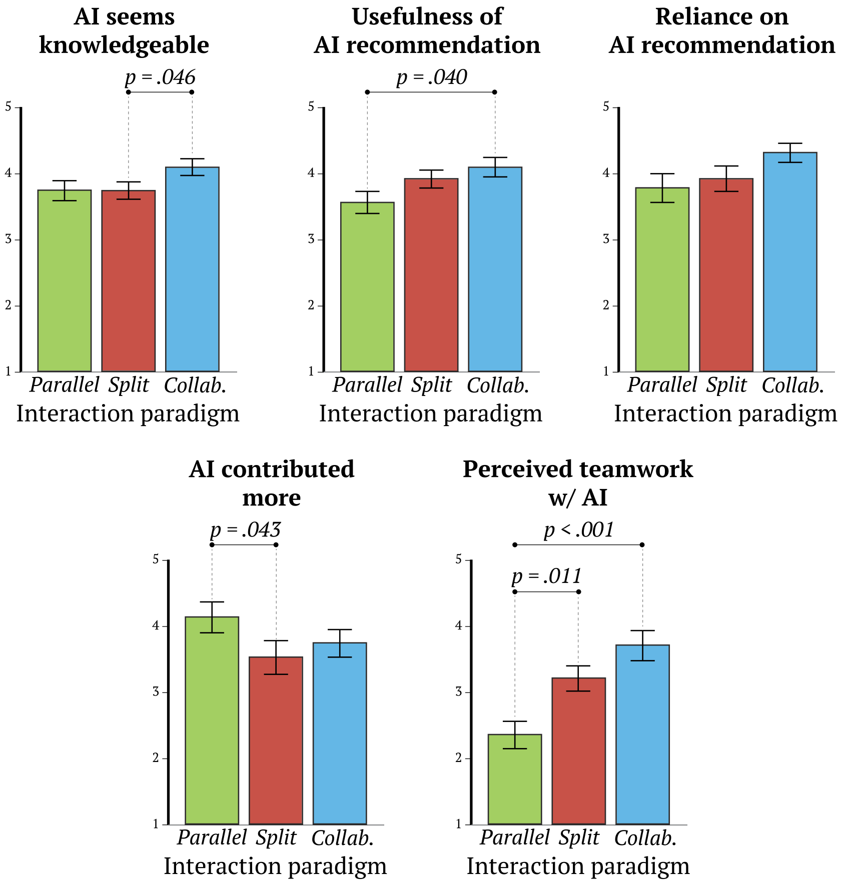


Fig. 7. Bar plots of the subjective metrics of user perception regarding the AI's knowledge, the usefulness of and user reliance on its recommendations, its contribution to the task, and sense of user–AI teamwork. Significant comparisons are displayed. Error bars correspond to standard errors.

4.7. Does the level of user involvement affect non-expert users' perceptions of an AI agent when interacting with it?

We conducted a one-way repeated measures ANOVA for each subjective questionnaire item separately where the interaction paradigm was set as a fixed effect and participants as a random effect. We describe the results for the questions for which we observed significant differences in the following subsections and Fig. 7 presents the ratings for these questions; Table A.1 in the Appendix summarizes the statistics and test results for all of the questionnaire items. Overall, most participants (75%) preferred the CDM interaction paradigm for future interactions, followed by the PDM (14%) and SDM (11%) paradigms.

4.7.1. Perceived knowledge of the AI agent

We found significant differences across interaction paradigms for users' ratings of the AI's perceived knowledge of the task ($F(2,54) = 4.12, p = .022, \eta_p^2 = .132$). Post-hoc pairwise comparisons using the Bonferroni correction revealed that users perceived on average the AI to be significantly more knowledgeable in the CDM than in the SDM interaction paradigm ($p = .046$).

4.7.2. Usefulness of the AI agent's suggestions

We found significant differences across interaction paradigms for users' ratings of the usefulness of the AI's suggestions ($F(2,54) = 4.72, p = .013, \eta_p^2 = .149$). Post-hoc pairwise comparisons using the Bonferroni correction revealed that users' average rating of the agent's usefulness was significantly higher in the CDM interaction paradigm than in the PDM (p = .040) paradigm, while no significant differences were observed against the SDM paradigm.

4.7.3. Reliance on the AI agent to complete the task

We found significant differences across interaction paradigms for users' ratings of their reliance on the AI to complete the task ($F(2,54) = 3.32, p = .044, \eta_p^2 = .110$). However, post-hoc pairwise comparisons using the Bonferroni correction did not reveal any significant difference on users' reliance ratings under different interaction paradigms.

4.7.4. Perceived contribution of the AI agent

We found significant differences across interaction paradigms for users' ratings regarding the AI's contribution to the task ($F(2,54) = 3.26, p = .046, \eta_p^2 = .108$). Post-hoc pairwise comparisons using the Bonferroni correction revealed that users perceived that, on average, the AI agent contributed more to the task than they did in the PDM interaction paradigm compared with the SDM ($p = .043$), but no significant differences were observed against the CDM paradigm.

4.7.5. Perception of teamwork in completing the task

We found significant differences across interaction paradigms for users' ratings on working together with the AI agent ($F(2,54) = 14.96, p < .001, \eta_p^2 = .356$). Post-hoc pairwise comparisons using the Bonferroni correction revealed that users' teamwork ratings were significantly higher in both the CDM ($M = 3.71, SD = 1.21, p < .001$) and SDM ($M = 3.21, SD = 0.99, p = .011$) interaction paradigms than in the PDM ($M = 2.36, SD = 1.10$) paradigm.

5. Discussion

We studied how interaction paradigms with different levels of user involvement affected people's behavior toward and perceptions of an AI agent in a task with a high knowledge imbalance. Our main findings suggest that users are more willing to agree with an AI's recommendations when they are more involved in its prediction generation process. In addition, undertrust outcomes were less frequent and users' positive perceptions of the AI agent and their collaboration as a team were superior in the more interactive paradigms than in the standard parallel AI-assisted decision-making condition.

5.1. Participants' willingness to follow the AI's recommendations

When analyzing participants' responses to whether or not they agreed with the bird categories suggested by the AI agent, our results indicate that answering *No* was hardly ever observed (5/672 trials), while the other two options (*Yes* and *Not sure*) were more commonly chosen in all three interaction paradigms. We attribute the lack of outright AI recommendation rejections to the knowledge imbalance in the task and participants' high rating of their familiarity with AI technologies; this self-rating could have biased users to inherently trust this specific agent given AI's general success in visual recognition tasks (Russakovsky et al., 2015), rather than showing algorithmic aversion (where we would have observed this same trend for the *Yes* answer).

Furthermore, considering the knowledge imbalance present in the task, it was difficult for users to judge the AI model's performance even after observing the model in practice. The AI's performance was far from accurate (its accuracy was around 63%), but this was never shown or stated to the participants during the task. While both stated and observed performance have been demonstrated to affect users' reliance on an AI partner (Yin et al., 2019), how people rely on an AI agent in a knowledge imbalance scenario should be further explored, especially under different accuracy levels as these can influence the development of trust (Papenmeier et al., 2022).

The interaction paradigms in which users played a more active role resulted in a higher likelihood that they agreed with and followed the recommendations presented by the AI agent compared to the standard PDM paradigm, providing support for H1. We attribute users' willingness to follow the AI's recommendations to the fact that they were part of its process in generating these predictions, which is typically obscured otherwise (i.e., "a black box") and can result in user distrust (Schaffer et al., 2019).

An alternative interpretation is that, when users were presented only with the AI's final bird category recommendations in the PDM interaction paradigm, they were less likely to follow those recommendations, which is supported by the majority of *Not sure* responses observed in this paradigm—that is, participants were less likely to answer *Yes* and follow the AI's recommendations. This finding is consistent with prior research demonstrating that novice users followed less frequently the recommendations from a diagnostic support tool that were presented as a direct or indirect cue to solve the main task (Chavaillaz et al., 2019).

However, blind agreement with an AI is not always desired, which is why we further explored overtrust and undertrust, noting that the agreement distribution determines desired and undesired trust outcomes. We observed more overtrust outcomes in the CDM interaction paradigms than in the standard PDM paradigm, signifying that participants were less likely to follow incorrect recommendations from the agent in the standard paradigm. We attribute this to the fact that, in the PDM paradigm, participants were more likely to answer *Not sure* when asked if they agreed with the AI's recommendations and were overall less likely to follow them—which, if incorrect, would have resulted in overtrust. Users' lack of involvement and knowledge of the inner workings of the AI model might have resulted in more caution and reluctance to blindly follow the AI's recommendations, reducing the possibility of participants experiencing overtrust. Although we did not observe a tendency to accept incorrect AI advice in the PDM interaction paradigm, previous work has reported that users with less expertise are prone to accept AI advice when it is incorrect (Micocci et al., 2021), as we observed in the CDM interaction paradigm. Users' involvement and closer collaboration with the AI agent promoted agreement with the AI's recommendations and approximates people's trust in terms of a behavioral measure (Yin et al., 2019). However, as people feel they can trust the AI more because they received assistance that they could understand (in form of the bird attributes) and apparently guided the model's outcome generation, they might end up following incorrect recommendations in the main task as well. In such case, the initial trust in the AI system became overtrust.

Meanwhile, we observed fewer undertrust outcomes in both the SDM and CDM interaction paradigms than in the standard PDM. In the former interaction paradigms, participants were more likely to agree with the AI's recommendations and less likely to reject them, which reduced the possibility of undertrusting then model when it was correct. These findings suggest that, despite the knowledge imbalance present in the main task, allowing users to provide information for the model to use in its prediction generation process encourages them to follow the AI's recommendations and thus avoid undertrust outcomes. Conversely, participants in the PDM interaction were more likely to undertrust the AI agent by rejecting its recommendations even when they might be correct; the rejection of the AI's recommendations is consistent with the majority of *Not sure* responses recorded in this interaction paradigm. These findings provide partial evidence for H2; while undertrust was less common in the two more involved interaction paradigms, overtrust was only reduced in the interaction paradigm with the least user involvement.

5.2. Additional behavioral indicators of trust

We further explored indirect indicators of trust by analyzing the attributes provided by participants when describing the bird images. When the participants were presented with recommendations on which characteristics to describe in the CDM interaction paradigm, the average number of attributes described was fewer than when the users did not have any guidance, as in the SDM interaction paradigm. We attribute this difference to the fact that in the former condition, participants could be more confident when the AI agent displayed its bird category recommendation because the agent itself had suggested which attributes were most important to provide. Furthermore, in the CDM paradigm, we analyzed the average percentage of attributes suggested by the AI agent that were present in participants' final bird descriptions with respect to the total description length; the high percentage of suggested attributes (84%) in the final descriptions could be an indicator of users' trust in the agent. In contrast, in the SDM interaction paradigm, users did not receive any information on which attributes were useful for the AI's recommendation process. Without this further guidance, their descriptions included on average more attributes. A potential explanation for longer bird descriptions is that users could keep adding more attributes either because they expected an updated recommendation with the additional attributes or because they thought that the AI agent might need more information to generate the bird category, given that they were not aware of the usefulness of the attributes they had provided for the AI model to distinguish the bird. However, we note that fixing the AI's bird category suggestions irrespective of user-provided attributes and minimum description length may have negatively affected users' perception of whether the AI had actually considered their input. This limitation could have ultimately affected users' willingness to provide more attributes for the AI model in both the SDM and CDM interaction paradigms.

Initially, we formulated the description length as a proxy for trust in the AI's recommendations. From users' behavior toward AI's recommendations, we found that in both the SDM and CDM paradigms users were more likely to follow the AI's bird category recommendations than in the PDM. However, our initial claim is not fully validated when considering the subjective assessment of trust. A correlation analysis between perceived levels of trust with the length of the descriptions for each interaction paradigm did not reveal a significant relationship between the two measures (SDM: $r(28) = 0.34, p = .074$ and CDM: $r(28) = 0.30, p = .122$). Furthermore, we did not observe significant differences in subjective trust between the two interaction paradigms that required participants to describe the birds. Therefore, these findings do not support H3. The concept of trust is challenging to capture in human–AI interactions as previous works have referred to the nuanced differences between attitudinal measures of trust and behavioral measures of reliance (Scharowski et al., 2022; Papenmeier et al., 2022).

As another indicator of user trust, we used the measurement of decision time when agreeing or disagreeing with the AI agent's recommendations. Considering participants' increased involvement during the main task, we found that having an active role in collaborating with the AI agent (i.e., in the SDM and CDM paradigms) resulted in users responding faster to the agreement question posed to them than in the PDM interaction paradigm. In the PDM condition, users had a passive role and may have needed additional time to compare the bird category suggested by the AI agent with the image presented to them. Meanwhile, in the SDM and CDM interaction paradigms, asking users to provide task-relevant information regarding the description of the bird required them to make an initial effort by looking at the finer details of the images in the first place. Additionally, since participants provided the input for the AI agent to generate its bird category suggestion, they may have been less hesitant about the final recommendation. However, participants' increased willingness to follow the AI's bird category recommendations in the more active paradigms (SDM and CDM) than in the standard PDM and faster responses in the active paradigms do not necessarily indicate higher levels of trust; a correlation analysis between perceived levels of trust with decision times did not reveal a significant relationship between the two measures ($r(84) = 0.07, p =$ .544). Therefore we did not find evidence for H3. Lastly, even though users could change their responses to the agreement question after seeing the AI's recommendation, we only observed such changes in a few trials (7/672), which we attribute to an overall lack of confidence the participants may have had due to the high knowledge imbalance.

5.3. Perceptions of the AI agent

Overall, users' reliance on the AI agent's recommendations and their perception of their usefulness, the agent's task knowledge, teamwork, and team members' contributions varied across the interaction paradigms. These concepts are related to the perceived capabilities of the AI agent and should be enhanced appropriately in successful interactions.

Participants' perceived usefulness and reliance ratings did not differ across the SDM and CDM paradigms, presumably because in the former, their perceptions still benefited from being involved in the AI's overall recommendation generation process. However, displaying updated attribute recommendations and requesting attributes to describe resulted in higher usefulness ratings than in the PDM paradigm, where users interacted with a black box AI agent. The task-relevant suggestions on bird attributes to provide in the CDM interaction paradigm could be easily understood by users, contributing to a better perception of the usefulness of the AI's suggestions to complete the main decision-making task.

The AI agent provided bird category suggestions in all three interaction paradigms with the same error rate, but users perceived the AI to be more knowledgeable when they were further presented with additional suggestions on bird attributes to describe in the CDM interaction paradigm than when the AI agent only provided a bird category suggestion following the user's input in the SDM paradigm. We attribute participants' higher ratings of the AI's knowledge in the interaction paradigm where they collaborated more closely with the AI agent to generate the bird category recommendation to the fact that the AI further demonstrated knowledge of the relevant attributes to classify the bird in the image. However, no differences were found with respect to the PDM paradigm where the AI agent solely provided bird category recommendations. Even though in both the SDM and CDM paradigms users provided feature-level information that they believed was processed by the AI agent to update the bird category outcomes, the presence of feature-level feedback in a binary text classification task negatively affected perceived model accuracy (Smith-Renner et al., 2020); we hypothesize that this difference in user perception is due to the knowledge imbalance in our task, which made it difficult for users to detect the model's weaknesses or flaws.

The level of user involvement in the studied interaction paradigms determined the roles and contributions of both the advisor and the advisee. By giving both the AI agent and the user the opportunity to contribute to the task's completion under a knowledge imbalance condition, participants had a better impression of working together with the AI when they collaborated by providing descriptive bird attributes—either in the SDM or CDM interaction paradigms—than when they merely accepted or rejected recommendations in the PDM paradigm. We were further interested in users' perceptions of the AI agent's contribution to a task under a knowledge imbalance condition. In the PDM paradigm, in which users were not involved in the AI's process of generating a bird category recommendation, the AI's contribution was perceived to be greater as compared to the SDM paradigm, where users alone guided the AI's prediction generation process. Giving users some sense of control over the AI's recommendations made them feel that they contributed more to complete the task. However, we did not observe such a difference in the perceived contribution of the AI agent when comparing against the CDM interaction paradigm, in which the AI agent also made recommendations on what task-relevant information to provide as users similarly guided the prediction generation process. Overall, users' perceptions of the AI agent were improved in the interaction paradigms with more user involvement, providing evidence for H4. Furthermore, participants may have potentially preferred the AI agent in the CDM interaction paradigm only due to their own contributions and the additional guidance provided by the AI agent allowing them to build a prediction together as a team.

5.4. Implications for designing AI-assisted decision-making systems that account for high knowledge imbalance

The design of human–AI interactions should consider both the end users of the AI system and their contextual knowledge, accounting for potential differences in domain knowledge for diverse reasons. In our study, we contextualized such a knowledge gap between an AI agent and its users in a task wherein our participants were all novices and therefore could not complete said task on their own. One of the main challenges in achieving effective collaboration with an AI agent under a knowledge imbalance condition is a novice's limited ability to appropriately calibrate their trust (Gaube et al., 2021; Nourani et al., 2020), whereas experts can use their domain knowledge to evaluate AI recommendations more critically, question those recommendations' validity, and better identify the limitations of the AI agent. In an attempt to allow novice users to gain insight into the task at hand, we provided them with the opportunity to guide the AI's prediction generation process via task information that they were familiar with—in our task setting, a description of visual information. We identified potential benefits to this paradigm, such as better perception of the AI agent and increased willingness to follow its recommendations, which ultimately aids in trust development; however, trust calibration outcomes should be further adjusted to avoid undesired behaviors such as overtrusting the AI agent. In our study design, we considered users' opportunities to provide input to the model in isolation, but such interactions can be complemented with additional insights from the AI model, such as explanations (Smith-Renner et al., 2020) or providing uncertainty measures of its suggestions. Furthermore, designers should consider the trade-off between designing for deeper engagement in human–AI interactions and the efficiency and usability of the AI system (Gajos and Mamykina, 2022) depending on the task domain. We recommend that future AI system designers incorporate features in the joint decision-making process that allow users to gain further insight into the task at hand so as to better validate the AI's recommendations.

5.5. Limitations and future work

Our results may reflect the perceptions of a specific group of people, as the participants in our study were mostly recruited from close contacts in an academic environment in which people were abnormally familiar with AI techniques, as confirmed by the high self-rating of familiarity with AI technologies (4.12 ($SD = 0.88$)). Further empirical validations with more diverse and larger groups of users can provide additional insights on the effects of different interaction paradigms when working with AI systems.

Based on users' feedback and the distribution of user responses, we recognize that the knowledge gap present in this task might have been considerably large, resulting in users with zero knowledge of how to complete the task. Therefore, even if users did not respond randomly to the agreement question at the end of each categorization trial, it would still be difficult to tell if they were guessing when considering each AI suggestion. Furthermore, the fixed, static predictions of the AI agent were a simplification for our implementation, but we acknowledge that users' involvement and collaboration with the AI agent could be improved with a more dynamic model. Similarly, as no feedback from the AI system was presented during the experiment, it would have been difficult for participants without previous task knowledge to develop an accurate mental model of the AI.

For future work, we recommend testing different levels of user involvement in tasks with lesser knowledge gaps to determine if the same benefits to teamwork perception are observed when users can be more independent and may not need the AI's assistance to complete the task. Besides, as AI systems will potentially assist non-expert users in certain circumstances, we encourage considering strategies that provide training or embed domain expertise for users to interpret and use the AI system appropriately.

6. Conclusion

Designing human–AI interactions is especially challenging when there is a knowledge imbalance present. Previous studies have demonstrated that the benefit from providing explanations is reduced when users lack domain expertise—as they cannot extract any meaningful insights (Wang and Yin, 2021)—and that providing additional information can be misleading for users even in simple tasks (Suresh et al., 2020) or may fail to increase trust or performance (Cheng et al., 2019; Nourani et al., 2021). Our findings suggest that involving users in the decision-making process by giving them an active role has the potential to enhance user perception of an AI system, but appropriately calibrating their trust for successful teaming outcomes remains a challenge.

CRediT authorship contribution statement

Catalina Gomez: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. **Mathias Unberath:** Conceptualization, Methodology, Writing – original draft, Supervision, Project administration. **Chien-Ming Huang:** Conceptualization, Methodology, Software, Formal analysis, Resources, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Funding

This work was in part supported by the National Science Foundation award #1840088.

Appendix

Subjective questionnaire. The list of the statements presented to participants read as follows:

- I trusted the AI's suggestions.
- The AI's suggestions were helpful.
- The AI seemed to be knowledgeable about the task.
- I relied on the AI's suggestions in completing the task.
- The AI's suggestions were questionable.
- The AI and I worked together as a team to complete the task.
- The AI contributed significantly to the completion of the task.
- The AI contributed to the task more than I did.
- I would not have been able to successfully complete the task without the AI.

Table A.1
Summary of users' ratings in response to statements in the subjective questionnaire across interaction paradigms. The scale of the means and standard deviations is 1 (Strongly disagree) to 5 (Strongly agree). $p < .05$ is considered a significant effect.

Question	PDM	SDM	CDM	F	p
Trust in the AI's suggestions	$M = 3.43$ $SD = 0.69$	$M = 3.64$ $SD = 0.68$	$M = 3.75$ $SD = 0.93$	$F(2, 54) = 2.67$	$p = .078$
Usefulness of the AI's suggestions	$M = 3.57$ $SD = 0.88$	$M = 3.93$ $SD = 0.72$	$M = 4.11$ $SD = 0.79$	$F(2, 54) = 4.72$	$p = .013$
AI's knowledge of the task	$M = 3.75$ $SD = 0.80$	$M = 3.75$ $SD = 0.70$	$M = 4.11$ $SD = 0.69$	$F(2, 54) = 4.12$	$p = .022$
Reliance on the AI's suggestions to complete the task	$M = 3.79$ $SD = 1.17$	$M = 3.93$ $SD = 1.02$	$M = 4.32$ $SD = 0.77$	$F(2, 54) = 3.32$	$p = .044$
The AI's suggestions were considered questionable	$M = 2.89$ $SD = 0.92$	$M = 2.71$ $SD = 0.94$	$M = 2.57$ $SD = 0.88$	$F(2, 54) = 1.56$	$p = .23$
Perceived teamwork	$M = 2.36$ $SD = 1.10$	$M = 3.21$ $SD = 0.99$	$M = 3.71$ $SD = 1.21$	$F(2, 54) = 14.96$	$p < .001$
The AI contributed significantly to the task	$M = 3.93$ $SD = 1.36$	$M = 4.00$ $SD = 1.19$	$M = 4.21$ $SD = 1.03$	$F(2, 54) = 0.834$	$p = .44$
The AI contributed more to the task	$M = 4.14$ $SD = 1.24$	$M = 3.54$ $SD = 1.35$	$M = 3.75$ $SD = 1.11$	$F(2, 54) = 3.26$	$p = .046$
Perceived inability to complete the task without the AI's assistance	$M = 4.11$ $SD = 1.13$	$M = 3.96$ $SD = 1.14$	$M = 4.32$ $SD = 0.98$	$F(2, 54) = 2.14$	$p = .14$

References

Amershi, Saleema, Cakmak, Maya, Knox, William Bradley, Kulesza, Todd, 2014. Power to the people: The role of humans in interactive machine learning. AI Mag. 35 (4), 105–120.

Angerschmid, Alessa, Zhou, Jianlong, Theuermann, Kevin, Chen, Fang, Holzinger, Andreas, 2022. Fairness and explanation in AI-informed decision making. Mach. Learn. Knowl. Extr. 4 (2), 556–579.

Bansal, Gagan, Nushi, Besmira, Kamar, Ece, Lasecki, Walter S., Weld, Daniel S., Horvitz, Eric, 2019a. Beyond accuracy: The role of mental models in human-AI team performance. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7. pp. 2–11.

Bansal, Gagan, Nushi, Besmira, Kamar, Ece, Weld, Daniel S., Lasecki, Walter S., Horvitz, Eric, 2019b. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. pp. 2429–2437.

Bansal, Gagan, Wu, Tongshuang, Zhou, Joyce, Fok, Raymond, Nushi, Besmira, Kamar, Ece, Ribeiro, Marco Tulio, Weld, Daniel, 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. pp. 1–16.

Berlyand, Yosef, Raja, Ali S., Dorner, Stephen C., Prabhakar, Anand M., Sonis, Jonathan D., Gottumukkala, Ravi V., Succi, Marc David, Yun, Brian J., 2018. How artificial intelligence could transform emergency department operations. Am. J. Emerg. Med. 36 (8), 1515–1517.

Buçinca, Zana, Malaya, Maja Barbara, Gajos, Krzysztof Z., 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 5 (CSCW1), 1–21.

Cai, Carrie J., Reif, Emily, Hegde, Narayan, Hipp, Jason, Kim, Been, Smilkov, Daniel, Wattenberg, Martin, Viegas, Fernanda, Corrado, Greg S., Stumpe, Martin C., et al., 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–14.

Chavaillaz, Alain, Schwaninger, Adrian, Michel, Stefan, Sauer, Juergen, 2019. Expertise, automation and trust in X-ray screening of cabin baggage. Front. Psychol. 10, 256.

Chen, Haomin, Gomez, Catalina, Huang, Chien-Ming, Unberath, Mathias, 2022. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit. Med. 5 (1), 1–15.

Chen, Haomin, Liu, T.Y., Correa, Zelia, Unberath, Mathias, 2020. An interactive approach to region of interest selection in cytologic analysis of uveal melanoma based on unsupervised clustering. In: International Workshop on Ophthalmic Medical Image Analysis. Springer, pp. 114–124.

Cheng, Hao-Fei, Wang, Ruotong, Zhang, Zheng, O'Connell, Fiona, Gray, Terrance, Harper, F. Maxwell, Zhu, Haiyi, 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. pp. 1–12.

Chiang, Chun-Wei, Yin, Ming, 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In: 13th ACM Web Science Conference 2021. pp. 120–129.

Cohen, Jacob, 1988. Statistical Power Analysis for the Behavioral Sciences. Routledge.

De-Arteaga, Maria, Fogliato, Riccardo, Chouldechova, Alexandra, 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Diao, James A., Wang, Jason K., Chui, Wan Fung, Mountain, Victoria, Gullapally, Sai Chowdary, Srinivasan, Ramprakash, Mitchell, Richard N., Glass, Benjamin, Hoffman, Sara, Rao, Sudha K., et al., 2021. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. Nature Commun. 12 (1), 1–15.

Dietvorst, Berkeley J., Simmons, Joseph P., Massey, Cade, 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Manage. Sci. 64 (3), 1155–1170.

Dodge, Jonathan, Liao, Q. Vera, Zhang, Yunfeng, Bellamy, Rachel K.E., Dugan, Casey, 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. pp. 275–285.

Eiband, Malin, Schneider, Hanna, Bilandzic, Mark, Fazekas-Con, Julian, Haug, Mareike, Hussmann, Heinrich, 2018. Bringing transparency design into practice. In: 23rd International Conference on Intelligent User Interfaces. pp. 211–223.

Gajos, Krzysztof Z., Mamykina, Lena, 2022. Do people engage cognitively with ai? Impact of AI assistance on incidental learning. In: 27th International Conference on Intelligent User Interfaces. pp. 794–806.

Gaube, Susanne, Suresh, Harini, Raue, Martina, Merritt, Alexander, Berkowitz, Seth J., Lermer, Eva, Coughlin, Joseph F., Guttag, John V., Colak, Errol, Ghassemi, Marzyeh, 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit. Med. 4 (1), 1–8.

Holzinger, Andreas, 2021. The next frontier: AI we can really trust. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 427–440.

Holzinger, Andreas, Dehmer, Matthias, Emmert-Streib, Frank, Cucchiara, Rita, Augenstein, Isabelle, Del Ser, Javier, Samek, Wojciech, Jurisica, Igor, Díaz-Rodríguez, Natalia, 2022. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf. Fusion 79, 263–278.

Holzinger, Andreas, Langs, Georg, Denk, Helmut, Zatloukal, Kurt, Müller, Heimo, 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 9 (4), e1312.

Kunkel, Johannes, Donkers, Tim, Michael, Lisa, Barbu, Catalin-Mihai, Ziegler, Jürgen, 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Lai, Vivian, Chen, Chacha, Liao, Q. Vera, Smith-Renner, Alison, Tan, Chenhao, 2021. Towards a science of human-AI decision making: A survey of empirical studies. arXiv preprint arXiv:2112.11471.

Lai, Vivian, Tan, Chenhao, 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. pp. 29–38.

Lazar, Jonathan, Feng, Jinjuan Heidi, Hochheiser, Harry, 2017. Research Methods in Human-Computer Interaction. Morgan Kaufmann.

Levy, Ariel, Agrawal, Monica, Satyanarayan, Arvind, Sontag, David, 2021. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–13.

Liao, Q. Vera, Gruen, Daniel, Miller, Sarah, 2020. Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–15.

Lima, Gabriel, Grgić-Hlača, Nina, Cha, Meeyoung, 2021. Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–17.

Lu, Zhuoran, Yin, Ming, 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–16.

Mahmood, Amama, Ajaykumar, Gopika, Huang, Chien-Ming, 2021. How mock model training enhances user perceptions of AI systems. arXiv preprint arXiv:2111.08830.

Micocci, Massimo, Borsci, Simone, Thakerar, Viral, Walne, Simon, Manshadi, Yasmine, Edridge, Finlay, Mullarkey, Daniel, Buckle, Peter, Hanna, George B., 2021. Do GPs trust artificial intelligence insights and what could this mean for patient care? A case study on GPs skin cancer diagnosis in the UK.

Miller, Tim, 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1–38.

Mohseni, Sina, Zarei, Niloofar, Ragan, Eric D., 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4), 1–45.

Ngo, Thao, Kunkel, Johannes, Ziegler, Jürgen, 2020. Exploring mental models for transparent and controllable recommender systems: a qualitative study. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. pp. 183–191.

Nourani, Mahsan, King, Joanie, Ragan, Eric, 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 8. pp. 112–121.

Nourani, Mahsan, Roy, Chiradeep, Block, Jeremy E., Honeycutt, Donald R., Rahman, Tahrima, Ragan, Eric, Gogate, Vibhav, 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In: 26th International Conference on Intelligent User Interfaces. pp. 340–350.

Papenmeier, Andrea, Kern, Dagmar, Englebienne, Gwenn, Seifert, Christin, 2022. It's complicated: The relationship between user trust, model accuracy and explanations in AI. ACM Trans. Comput.-Hum. Interact. 29 (4), 1–33.

Rastogi, Charvi, Zhang, Yunfeng, Wei, Dennis, Varshney, Kush R., Dhurandhar, Amit, Tomsett, Richard, 2022. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 6 (CSCW1).

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2016. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Schaffer, James, O'Donovan, John, Michaelis, James, Raglin, Adrienne, Höllerer, Tobias, 2019. I can do better than your AI: expertise and explanations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. pp. 240–251.

Scharowski, Nicolas, Perrig, Sebastian A.C., von Felten, Nick, Brühlmann, Florian, 2022. Trust and reliance in XAI–distinguishing between attitudinal and behavioral measures. arXiv preprint arXiv:2203.12318.

Schoonderwoerd, Tjeerd A.J., Jorritsma, Wiard, Neerincx, Mark A., Van Den Bosch, Karel, 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. Int. J. Hum.-Comput. Stud. 154, 102684.

Smith-Renner, Alison, Fan, Ron, Birchfield, Melissa, Wu, Tongshuang, Boyd-Graber, Jordan, Weld, Daniel S, Findlater, Leah, 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ML. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–13.

Suresh, Harini, Gomez, Steven R., Nam, Kevin K., Satyanarayan, Arvind, 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–16.

Suresh, Harini, Lao, Natalie, Liccardi, Ilaria, 2020. Misplaced trust: Measuring the interference of machine learning in human decision-making. In: 12th ACM Conference on Web Science. pp. 315–324.

Van Berkel, Niels, Goncalves, Jorge, Russo, Daniel, Hosio, Simo, Skov, Mikael B., 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–13.

Wall, Emily, Ghorashi, Soroush, Ramos, Gonzalo, 2019. Using expert patterns in assisted interactive machine learning: A study in machine teaching. In: IFIP Conference on Human-Computer Interaction. Springer, pp. 578–599.

Wang, Xinru, Yin, Ming, 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In: 26th International Conference on Intelligent User Interfaces. pp. 318–328.

Welinder, Peter, Branson, Steve, Mita, Takeshi, Wah, Catherine, Schroff, Florian, Belongie, Serge, Perona, Pietro, 2010. Caltech-UCSD birds 200.

Yang, Fumeng, Huang, Zhuanyi, Scholtz, Jean, Arendt, Dustin L, 2020. How do visual explanations foster end users' appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent User Interfaces. pp. 189–201.

Yin, Ming, Wortman Vaughan, Jennifer, Wallach, Hanna, 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. pp. 1–12.

Zhang, Qiaoning, Lee, Matthew L., Carter, Scott, 2022. You complete me: Human-AI teams and complementary expertise. In: CHI Conference on Human Factors in Computing Systems. pp. 1–28.

Zhang, Yunfeng, Liao, Q. Vera, Bellamy, Rachel K.E., 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. pp. 295–305.