Cognition, Technology & Work (2024) 26:435–455
https://doi.org/10.1007/s10111-024-00765-7

ORIGINAL ARTICLE

Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach

Allyson I. Hauptman[1] · Beau G. Schelble[1] · Wen Duan[1] · Christopher Flathmann[1] · Nathan J. McNeese[1]

Received: 4 December 2023 / Accepted: 17 April 2024 / Published online: 18 May 2024

Abstract
An obstacle to effective teaming between humans and AI is the agent's "black box" design. AI explanations have proven benefits, but few studies have explored the effects that explanations can have in a teaming environment with AI agents operating at heightened levels of autonomy. We conducted two complementary studies, an experiment and participatory design sessions, investigating the effect that varying levels of AI explainability and AI autonomy have on the participants' perceived trust and competence of an AI teammate to address this research gap. The results of the experiment were counter-intuitive, where the participants actually perceived the lower explainability agent as both more trustworthy and more competent. The participatory design sessions further revealed how a team's need to know influences when and what teammates need explained from AI teammates. Based on these findings, several design recommendations were developed for the HCI community to guide how AI teammates should share decision information with their human counterparts considering the careful balance between trust and competence in human-AI teams.

Keywords Human-AI teaming · Adaptive autonomy · Explainable AI · Artificial intelligence

1 Introduction

Modern advances in artificial intelligence (AI) continue to enable the creation of AI agents that can operate with increasingly higher levels of autonomy (LOA) (Chen et al. 2022). These higher LOA center around agents capable of performing tasks from start to finish with minimal human input and direct control (O'Neill et al. 2020; Parasuraman et al. 2000a), which enable AI agents to fulfill independent roles in a variety of teams, organizations, and task environments (McNeese et al. 2018; Wilson and Daugherty 2018). Consequentially, AI agents, in many situations, have become more than tools used *by* the team, but rather *part* of the team (O'Neill et al. 2020; McNeese et al. 2018). These new human-AI teams are able to leverage the technical strengths of AI and present humans and organizations with the ability to overcome existing struggles with all-human teams, such as operating in data-intensive and geographically distant contexts (Nyre-Yu et al. 2019; Chen 2023). While the unique information processing capabilities of AI make the prospect of these teammates new and exciting, their use also comes with unique challenges for teams.

AI agents capable of taking on independent team roles can operate with less human monitoring and control. Still, in complex environments involving elevated levels of uncertainty and risk, this lack of human oversight can lead to disastrous outcomes (Pedreschi et al. 2019; Suzanne Barber et al. 2000). This is because as systems execute decisions more independently, human situational awareness of the system's decisions decreases (Wickens et al. 2010). This issue is exacerbated by human distrust of AI systems that make decisions within a "black box" algorithm, which hides what and how the AI is processing information to make its decisions (Castelvecchi 2016). In response to this, methods for AI to provide explanations for their decisions have been developed

✉ Allyson I. Hauptman
ahauptm@clemson.edu

Beau G. Schelble
bschelble@clemson.edu

Wen Duan
wend@clemson.edu

Christopher Flathmann
cflathm@clemson.edu

Nathan J. McNeese
mcneese@clemson.edu

[1] School of Computing, Clemson University, 821 McMillan Rd., Clemson 29631, SC, USA

as one way to reduce the mystery of highly autonomous AI's "black box" decision-making nature (Shin 2021a; Weitz et al. 2019). However, there is a trade-off between too much and too little explanation (Dhanorkar et al. 2021). While explanations provide teammates with detailed information to better understand the rationale and intention behind the AI's decision, sometimes too much information can lead to cognitive overload and an inability for humans to focus on their own tasks, which can significantly frustrate a team's ability to work interdependently (Wang et al. 2019). Additionally, AI agent explanations must fit the communication needs of their human teammates (Stowers et al. 2021), which vary based explicitly on the team's working environment (Jarrahi et al. 2022). This means that the specific information that an AI agent communicates in its explanations is also extremely important to consider.

Previous research on AI autonomy has found that it would be beneficial if AI teammates were capable of operating at multiple levels of autonomy, based on changing tasks and environments (Hauptman et al. 2022; Zieba et al. 2010). The established benefits of dynamic autonomy levels raise the question of whether AI teammates should also possess different levels of explainability. There is already evidence to support the idea that explainability should not be a static feature, as human-computer interaction (HCI) research has found that AI needs to explain itself differently based upon what and to whom it is communicating (Dhanorkar et al. 2021). This is especially important for human-AI teams (HATs) because humans want the AI to adapt its interaction behaviors to be as helpful as it can be while keeping humans knowledgeable of essential information (Liao et al. 2020). In fact, research shows just the perception of AI as adaptive can increase human performance (Kosch et al. 2023). Despite robust research into how to make AI algorithms more transparent and explainable to the user (Larsson and Heintz 2020; Waltl and Vogl 2018; Hussain et al. 2021), there have been increased calls for more research into the content and frequency of explanations that humans need while interacting with an AI agent (Weber et al. 2015; Schoenherr et al. 2023).

Explainability and autonomy levels substantially contribute to trust development and growth in human-AI teams. The ability to understand an AI agent's capabilities and decisions is fundamental to a human's notion of its trustworthiness (Jacovi et al. 2021; Caldwell et al. 2022). This is because it allows them to predict the AI's future behavior (Jacovi et al. 2021). In fact, research into explainable agents in human–machine teaming has shown that explanations can substantially increase human teammate trust in the robot's decisions (Wang et al. 2016). Previous research on information needs shows that human interactions with technology affect the information they will perceive, accept, and trust from that technology, particularly in teams (Huvila et al. 2022). However, individuals' information needs may not be static and constant as they interact with technology. For instance, increasing familiarity with a specific technology eliminates the need to understand every detail of how it works (Hauptman et al. 2022). Additionally, the degree to which a human is "in the loop" of AI's decision-making process may fundamentally change how much and what information humans need to know and, in turn, change how they interact with and perceive the AI (Abbass 2019). Despite research into how AI explainability affects human behaviors, little is known with respect to the relationship between how much an AI teammate explains with how much autonomy it exhibits in executing its tasks. Lower autonomy systems must generally communicate more with humans due to the requirement for human input in their decisions. Thus, AI that provides a high or low level of explainability may also be *perceived* by a human teammate as even more or less autonomous. In order to investigate this relationship, this study explores the following research questions:

RQ1: How does teaming with an AI agent with a high or low level of explainability affect the human teammates' perceived trust and competence of the AI at both a low and high level of autonomy?

RQ2: How should the content of AI explanations change as the AI teammate's autonomy level changes?

Given the complex and context-dependent nature of teaming and explainability requirements, this research takes a mixed methods approach, utilizing two studies to answer the above research questions. In the first study, we conducted a 2x2 (LOA x Explainability Level) online networking experiment to examine the effects of different LOAs and AI explainability levels on participants' perceived trust and competency of their AI teammates. Then, in the second study, we held participatory design sessions with twelve of those participants in order to further understand the explainability needs and desires of human teammates for AI agents with varying LOAs. The identified dimensions of the dynamic relationship between the levels of autonomy and explainability of AI teammates are heavily grounded in both the participants' professional experiences and interactions with the AI in these studies. The resulting discussion and design recommendations provide an empirical starting point for the HCI community to model and understand the optimal explainability levels for AI teammates operating with different autonomy levels. This greatly contributes to the body of human-AI teaming literature as the community seeks to envision and design artificial agents that can work closely with and support humans in complex team environments.

2 Related work

In this section, we will lay the groundwork for our studies, beginning with the need for and types of AI explanations, followed by levels of AI autonomy. Finally, we will articulate the research gaps that motivate our research.

2.1 AI explanations

Previously, AI models have often been described as a black box into which information is simply input; the box "does its magic" and produces some form of output (Xu et al. 2019). Research has shown that these black-box models can have significant negative impacts when AI is used in complex situations (Cohen et al. 2021), such as the inability to track where something went wrong (Yu and Alì 2019). Some within the AI community have indicated a distinct lack of work into the ethics surrounding AI design (Slota et al. 2022). Cohen and colleagues found that minor mistakes in the training phase often led to severe issues with the model that could be relatively difficult to find and understand because of the model's lack of explanation (Cohen et al. 2021). Additionally, evaluations of medical AI technologies have demonstrated that black-box AI agents hinder their use and effectiveness due to ethical concerns (Duan et al. 2019). Opaque AI can have major negative implications for the humans with whom it interacts. For instance, research on AI-enabled recommender systems showed that opaque recommendations could decrease user self-confidence (Shin 2021a). In response to these challenges, a quickly growing area of research is ways to design AI to explain better reasoning and actions to humans (Xu et al. 2019; de Lemos and Grześ 2019; Pokam et al. 2019). User-centered explanation solutions attempt to alleviate these issues by developing AI that explains not only what it did but also why it did it in ways a human would understand (Wang et al. 2019). In regards to the *what*, the AI's output must be readable by the human audience, a concept often referred to as interpretability (Lipton 2018). Research shows that this interpretability encourages user trust in AI algorithms (Shin 2021b). As a function of that interpretability, the audience must be able to grasp what the output means, referred to as the agent's *understandability* (Joyce et al. 2023). Both of these aspects contribute to the delivery of an effective AI explanation (Marcinkevičs and Vogt 2020).

The research on explainable systems is exploding at such a rate that multiple reviews in the HCI (Speith 2022; Mueller et al. 2019) and computer science (Vilone and Longo 2020; Das and Rad 2020) communities have recently proposed new methods for organizing the subject. While these reviews focus widely on how the AI itself should be designed, they lack a human-centered approach to AI explanations. A recent study on the role of information exchange in designing explainable systems argued that the current trend towards using AI techniques to explain AI is insufficient, and the explanation recipients need to be more involved in how AI explanations are created and given (Xie et al. 2022). There are various reasons for this need, including the importance of effective human-centered AI explanations in building trust in AI algorithms and overcoming gaps in AI transparency (Shin 2021a). User-centered explanation solutions attempt to alleviate these issues by developing AI that explains not only what it did but also why it did it in ways a human would understand (Wang et al. 2019). In regards to the *what*, the AI must provide its output in a readable manner, a concept often referred to as *interpretability* (Lipton 2018). Research shows that this interpretability encourages user trust in AI algorithms (Shin 2021b). As a function of that interpretability, the audience must be able to grasp what the output means, referred to as the agent's *understandability* (Joyce et al. 2023). While these terms often overlap, interpretability refers to the AI's ability to explain an abstract concept, while understandability refers to the AI's ability to make it understandable to an end-user (Vilone and Longo 2020). Both of these aspects contribute to the delivery of an effective AI explanation (Marcinkevičs and Vogt 2020). This gap in considering how the explanations provided by an AI teammate are received by a human teammate is a driving motivation behind this research. This is why the AI explanations in the high explainability condition in the first study include what information the AI considered in accomplishing its task.

Explainability exists on a spectrum regarding the type and amount of explanations that the AI can provide. For instance, Dazeley and colleagues organized Levels of AI Explanation into a pyramid based on human psychological needs (Dazeley et al. 2021). Other researchers have classified an AI's level of explainability based upon the AI's algorithms and capabilities (Arrieta et al. 2020), (Sokol and Flach 2020). Most of these descriptions can fall into two main categories, low-level vs. high-level explainability models. Low-level XAI gives basic information about its decision, potentially displaying the algorithm(s) behind it or giving a brief description of what it is supposed to do or the results it found. High-level XAI gives more detailed explanations of the entire process, including their decision logic (Sanneman and Shah 2020; Miller 2019). This is arguably an essential step for an AI teammate because the degree to which humans understand an AI agent can greatly affect their acceptance and trust of it (Xu et al. 2019; Bansal et al. 2021). Some explainability research has articulated this as

the stakeholder variable, a concept stating that, because the goal of explanations is to satisfy the expectations and goals of a stakeholder, that stakeholder's perceptions of the explanations are important (Langer et al. 2021).

Explanations are only as valuable as they are understood and accepted by the persons receiving them (Dazeley et al. 2021). In approaching what may be considered a "low" vs. a "high" level of explainability, we turn to the point of view of research done by Lombrozo and colleagues, which suggests that humans perceive the level of explainability to be higher when the explanations communicate more events in the most coherent manner (Lombrozo 2006). This follows the research of Dazeley and colleagues, who found the more contextual information the explanations include, the higher value it is to the persons receiving the explanation (Dazeley et al. 2021). It also reflects human-AI research that shows increased acceptance of AI-generated communications that appear more human-like (Shin 2022). This implies that very high-level explanations from an AI teammate should be frequent, human-readable, and provided within the context of the team activity. This study utilizes those principles in designing high-level explanations for the AI teammate in the experiment.

The introduction of AI explanations directly addresses a variety of the damaging pitfalls brought about by the black-box nature of AI agents. Some of these pitfalls, as discussed above, lead directly to decreased trust in and acceptance of AI decisions (Zhou and Chen 2019). This is why understanding the effects and design implications of explainability in HATs is so important, as trust in AI's explanations is a key part of its acceptance by the humans with whom it interacts (Ehsan and Riedl 2019). Still, we cannot assume that increasing the level of explainability from a low-level to a high-level model will directly lead to increased trust and performance. In a study of human-agent teaming in Minecraft, Paleja and colleagues found that while AI teammate explainability led to greater situational awareness and increased performance for novices, it did not equate directly to increased performance for more experienced individuals (Paleja et al. 2021). In fact, when the AI's explanations evolved to include a full decision tree, the novice participants experienced cognitive overload (Paleja et al. 2021). The literature clearly shows a need to strike the right balance between an AI teammate's explanations and its human teammates' cognitive capabilities in order to promote intra-team trust and performance in HATs (Nakahashi and Yamada 2021), particularly in complex and high-risk environments (Ha et al. 2020).

2.2 Levels of autonomy

Addressing the various levels of autonomy (LOA) for AI in human-AI teaming is the final concept necessary to motivate the current research, as it coincides directly with the need for XAI. Artificial agents can be programmed to operate with different levels of autonomous behavior. In order to categorize these levels, autonomy researchers have adapted the LOA (Parasuraman et al. 2000b) into three categories of autonomy: no autonomy, partial agent autonomy, and high agent autonomy (O'Neill et al. 2020). AI that requires human input to perform any decision or action is not, actually, autonomy, according to the literature, as it performs no independent role (O'Neill et al. 2020). Agents with partial and high LOAs are capable of taking on independent functions that not only define their autonomous behavior but also make them capable of taking on independent team roles (O'Neill et al. 2020), making them inherently more integral to the team than a simple tool.

Teams operate in dynamic, complex environments that change over time, and AI teammates need to be able to change their behavior and capabilities to match such changes (Suzanne Barber et al. 2000). This might mean that AI teammates may need to change their LOA over time, a concept known as adaptive autonomy (McGee and McGregor 2016). Furthermore, teammates not only need to adapt to their environments but also to their human teammates (McNeese et al. 2018; Richards and Stedmon 2017). This concept heavily motivates these studies' inquiries into the intersection of autonomy and explainability. If AI teammates need to adapt their autonomy levels in order to fulfill their team role while simultaneously adapting to the needs of their human teammates, then the explanations they provide to those human teammates may need likewise to adapt as their autonomy levels change.

Our review of the existing literature on autonomy and explainability levels in human-AI teaming presents a couple of intriguing research gaps. Previous work indicates that the black-box design of AI agents frustrates the human ability to understand and trust in an AI teammate's decisions (von Eschenbach 2021), but to what degree that frustration varies as AI autonomy varies is uncertain. Additionally, despite this recorded frustration, there is also evidence that higher-level explainability models also come with negative consequences and do not always lead to increased trust and performance (Paleja et al. 2021). To address these gaps, we designed two complementary studies that jointly provide a systematic understanding of the relationship between humans' nuanced needs for explainability and their AI teammate's level of autonomy.

3 Study 1

3.1 Methods

The experiment conducted for Study 1 specifically explores the effect of AI autonomy level and explainability on humans' perceptions of competence and trust in their AI teammates. Study 1 utilized a mixed 2 (AI Autonomy Level: Low, High) x 2 (AI Explainability Level: Low, High) experimental design, with the autonomy level of the AI agent manipulated between-subjects and the AI explainability level manipulated within-subjects. Participants teamed up with a single AI teammate to complete the Cisco network simulation program Packet Tracer. These human-AI team dyads completed two iterations of the Packet Tracer activity (described below). In the following section, we will overview the procedures for developing and implementing the experimental platform and performing the experiment with the participants.

3.1.1 Networking task

Study 1 used Cisco's educational networking simulation program, Packet Tracer, one of the most widely used visual learning methods for computer networking (Janitor et al. 2010). This program permits users to simulate the physical cabling of networking devices and the software configuration of the devices, making it ideal for the current study, as it allows for multiple networking tasks that could be performed by both the human and the AI team members simultaneously. It also showcased a very realistic human-AI teaming scenario, where an AI agent can quickly execute computer commands while a human accomplishes the physical tasks of which an AI agent is incapable. All four conditions of the task were selected for beginner-level participants, such that they would all be equally challenging and time-consuming for participants. Packet Tracer is also a lightweight program, meaning we could place the program on a virtual machine that our participants could log into from anywhere in the world. Screenshots of the virtual platform participants engaged with are shown in Fig. 1.

All four tasks focused on the setup and configuration of a small-scale local network. During the experiment, participants played the part of the Physical Network Tech, responsible for powering the devices and moving physical cables, which in Packet Tracer equates to the participant dragging and dropping the cables between the correct devices. Meanwhile, the AI agent played the role of Software Tech, responsible for all the actual device configurations. This job role also helped minimize the effect of lower experience levels, as the tasks participants needed to perform were relatively simple and easy to learn through a short practice exercise. Pilot sessions for this study showed that a practice exercise prior to the start of the actual experiment was indeed extremely helpful to inexperienced participants, and so all participants did a practice exercise with the first author at

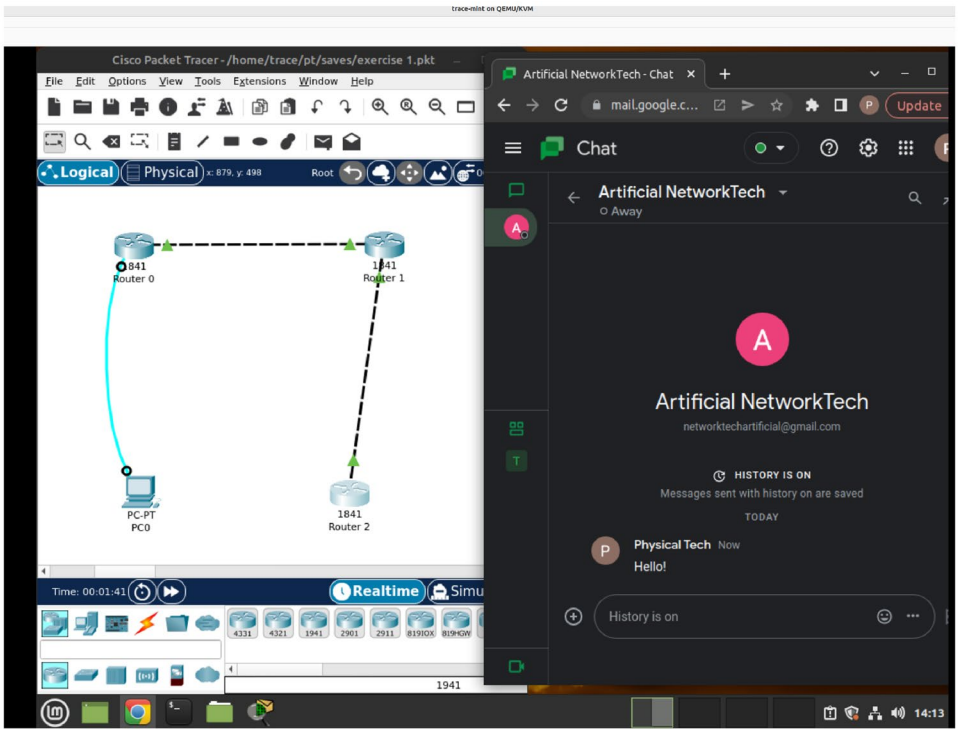

Fig. 1 Screenshot of the experiment platform. Participants are presented with the proper network devices and cables and are responsible for selecting and moving the blue console cable between devices for the AI to access the right device

the start of their virtual session to ensure they understood their tasks and how to communicate with their AI teammate.

3.1.2 AI teammate

The AI teammate implemented in Study 1 utilized the Wizard of Oz methodology (WoZ), a common technique within the HCI community (Kelley 2018). This technique enables researchers to simulate more advanced design features like AI teammate communication to garner insights regarding AI teammates of the future. The virtual platform further supported this technique, as the participants did not know that the chats they had with the AI were actually being conducted by a confederate researcher following a pre-made script developed throughout several piloting sessions to ensure accuracy and applicability. All communication with the participants from the researchers occurred using a separate chat to maintain the script.

Between-subjects manipulation: autonomy level Exit interviews from the pilot sessions showed that because the AI exists only as a chat agent, an explicit permission phrase was the best option to effectively delineate autonomy levels to participants. Specifically, in the "Low Autonomy" condition, the AI teammate had the confederate ask permission for all actions taken during the exercises and could not move forward in the task without the participant granting that permission. Alternatively, in the "High Autonomy" condition, the AI teammate had the confederate inform the participant what the AI would do but did not ask or require their permission to perform the action. For both conditions, the confederate had a set of predetermined responses to any questions posed to the agent by a participant that was in line with the AI agent's supposed autonomy level.

Within-subjects manipulation: explainability Finally, the pilot sessions also informed the design of the AI explainability manipulations by tying it to the Packet Tracer task. In particular, pilot participants wanted the explanations to be contextually tied to the interface. As such, the "High Explainability" AI teammate was defined by the AI opening the console with the commands it used to program the networking device and explaining why it used the commands. Whereas the "Low Explainability" AI was distinguished by the AI simply telling the participant when it was starting and completing a task. These are apparent changes in the amount of information that the AI teammate was providing to the participant in terms of both content and frequency, which was deemed necessary after exit interviews from the pilot sessions that indicated the need for additional information in the "High Explainability" condition in the form of the console display being included for these participants to perceive the manipulation as expected. This visual form of explanation has been shown to allow participants to better calibrate trust in AI (Liu et al. 2023).

Table 1 Participant demographic information

Participants: 44 (M_{Age} = 34.63)				
Men	Women	Non-binary/ Third gender	Other	
25	19	0	0	
Caucasian	Black/African-American	American-Indian/ Alaska Native	Asian	**Other**
31	4	1	6	2
At least some information technology/Networking experience				
37				

3.1.3 Participants

Following approval from the Clemson University Institutional Review Board, the current study recruited 44 participants, with 19 identifying as women and the rest identifying as men. The average age of participants was 34.63 (for additional demographic information, see Table 1). Based on an a priori power analysis with an effect size of 0.13, in order to meet power, this experiment required a minimum sample size of 42 total participants, which was achieved. Participants were recruited using email solicitation and snowball sampling of individuals with experience in information technology and/or computer science disciplines. This inclusion criterion was implemented to help control for the potential confound of subject matter expertise by recruiting participants with generally equal levels of knowledge and aptitude for the Packet Tracer task, which was achieved as shown in Table 1. However, significant experience in networking work was not an explicit inclusion criterion, as the Packet Tracer task included training on the specific topics necessary to complete the task successfully and was designed for beginner knowledge levels.

3.1.4 Procedure

When participants agreed to participate in the experiment, they received an email with the task and descriptions of the exercises they would perform. They also received instructions for logging into the virtual machine using Chrome Remote Desktop. In case the participant was not familiar with the Packet Tracer program, a tutorial video was also included. Five minutes before their designated time, they received the access code for the machine and the link to the survey. Prior to beginning the experiment, participants completed the pre-task survey, which covered informed consent and demographic information.

After completing the pre-task survey, participants went on to complete a training period. The training period included written instructions with illustrations that described the task

and how to complete it with their AI teammate, defining the two roles and their interdependencies. This written training period was followed up with a live training phase where participants engaged in a live practice round of the task with their AI teammate and the ability to ask questions with the researcher should they have any. Once the training session was completed, the participants were randomly assigned to one of the two between-subjects conditions of either low or high AI teammate autonomy. All participants performed two exercises, one with an AI agent with High Explainability and one with an agent with Low Explainability. Half of the participants received the High Explainability condition first, and half received the Low Explainability condition first. This counter-balancing minimized the effect of participants having increased comfort and understanding with the exercises in the second condition they received and helped mitigate any potential spill-over effects between within-subjects conditions.

Following this exercise, participants began interfacing with their AI teammate for the first exercise. Upon the completion of the first exercise, participants were prompted to complete the next part of the survey, which measured their perceptions of the agent as trustworthy and competent for that exercise. Once complete, they conducted the second exercise, followed by the remaining portions of the survey. Once the survey was complete, the remote session was terminated. Participants received a debrief message following the completion of the survey explaining the intent of the exercise, the use of the WoZ setup, and thanking them for their time.

3.1.5 Measures

Human-AI interaction research has shown that the perceptions that users have of an AI agent are vital to the design of explainable systems, as these perceptions directly affect the user's acceptance and trust in the agent (Shin 2020). Thus, the main measures utilized in this study were self-reported perceptions of the participants on 5-item Likert scales, as described in the following paragraphs.

Trust in the agent Human trust in their AI teammates is integral to both their acceptance of the AI and the team's overall performance (Costa et al. 2018; Centeio Jorge et al. 2022). For this reason, trust was measured in the post-task survey after each interaction exercise with the AI agent using a 3-item 5-point Likert scale (1=strongly disagree, 5=strongly agree). Items included "The autonomous agent I worked with was trustworthy," "The autonomous agent I worked with could be trusted to complete the assigned tasks," and "I did not feel the need to monitor the autonomous agent's actions" and had a reliability of $\alpha = .80$. These questions were based on the outcomes of trust defined by Lumineau (Lumineau 2017) and adapted from previous use in human-AI teaming research (Schelble et al. 2022a, b).

Perceived competency of the agent Perceived Competency was measured in the post-task survey after each interaction exercise with the AI agent using a 3-item 5-point Likert scale (1=strongly disagree, 5=strongly agree) adapted by the authors based on similar perception of competence scales utilized in AI research Gieselmann and Sassenberg (2023). Items included "The autonomous agent I worked with was competent at its role," "The autonomous agent I worked with was capable of completing its assigned tasks," "The autonomous agent I worked with was capable of joint problem solving" and had a reliability of $\alpha = .79$.

Perceived awareness of the AI's actions Situation awareness is an important human factor for human-centered AI design (Chignell et al. 2023) and needed to be included in some fashion. We based this perceptual awareness measure on Tier 1 of the three-tier situational awareness model, where an entity must accurately perceive their surroundings (Endsley 1995). Thus, perceived situational awareness was measured in the post-task survey after each interaction exercise with the AI agent using a 1-item 5-point Likert scale (1=strongly disagree, 5=strongly agree) developed by the authors. The item stated, "I felt aware of the actions my autonomous teammate was taking."

Understanding of the AI's actions The second tier in the three-tiered model of situational awareness addresses an entity's ability to make sense of what they perceive, or *understand* their surroundings (Endsley 1995). Thus, we measured the participant's understanding of the AI's actions in the post-task survey after each interaction exercise with the AI agent using a 1-item 5-point Likert scale (1=strongly disagree, 5=strongly agree) developed by the authors. The item stated, "I understood why my autonomous teammate took certain actions." The final element of the three-tiered situational awareness model was explored in the participatory design sessions in Study 2.

3.2 Results

To address the stated research questions, a series of 2 (Level of Autonomy: Low, High) x 2 (AI Explainability: Low, High) mixed model ANOVAs were conducted on participants' survey responses after each teaming experience. The level of autonomy factor was conducted between-subjects, while AI explainability was analyzed as a within-subjects factor. The following sub-section reviews analyses on trust, perceived competence, perceived awareness, and understanding of the AI teammate, concluding with an analysis of the chat data to reveal participants' objective need for AI explainability during the task. The following results address RQ1, which sought to investigate how increases in

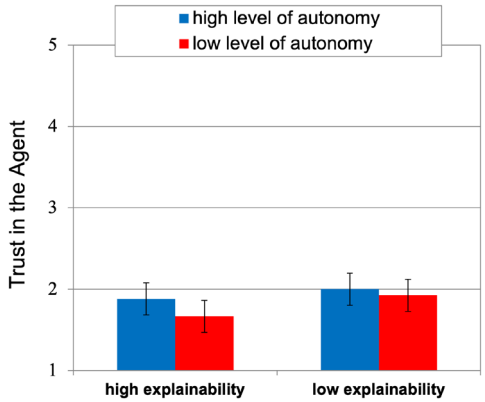

Fig. 2 Trust in the AI on a scale of 1 to 5, by the level of autonomy for high and low explainability (error bars represent standard error of the mean)

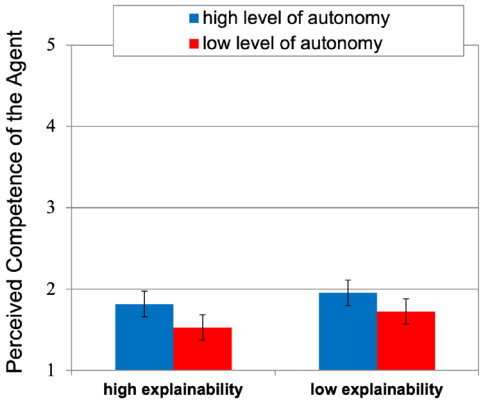

Fig. 3 Perceived competency of the AI on a scale of 1 to 5, by the level of autonomy for high and low explainability (error bars represent standard error of the mean)

the information given by AI explanations change humans' perception of their AI teammates at varying LOA.

3.2.1 Trust in the AI teammate

The main effect of AI teammate autonomy level on trust in the AI teammate was non-significant ($F(1, 42) = 0.30$, $p = 0.59$, $\eta^2 = 0.01$). However, the main effect of AI teammate explainability on trust in the AI was significant ($F(1, 42) = 4.42$, $p = 0.04$, $\eta^2 = 0.10$; see Fig. 2) and this was a medium-sized effect (Cohen 1988). Specifically, participants trusted the high-explainability AI teammate less ($M = 1.77$, $SE = 0.14$) than they trusted the low-explainability AI teammate ($M = 1.96$, $SE = 0.14$). Lastly, the interaction effect between autonomy and explainability levels was non-significant ($F(1, 42) = 0.57$, $p = 0.45$, $\eta^2 = 0.01$).

This result shows that the participants felt the AI teammate that explained all of the actions it took in completing its team tasks through the chat was less trustworthy than the AI that only told them when it was starting and completing a task. This result suggests that in some teams the additional communications from the AI teammate, possibly due to communication overload, is actually counterproductive to building trust. This result indicates that humans working in human-AI teams want information from their AI teammates only at appropriate intervals. In this case, this additional information hurt the participant's trust in the AI when it came during the task itself.

3.2.2 Perceived competence of the AI teammate

There was no significant main effect of AI teammate autonomy level on participants' perceived competence of the AI ($F(1, 42) = 1.50$, $p = 0.23$, $\eta^2 < 0.01$). However, there was a significant main effect of AI teammate explainability level on perceived competence ($F(1, 42) = 5.73$, $p = 0.02$, $\eta^2 = 0.12$; see Fig. 3), and this was a medium-sized effect (Cohen 1988). Specifically, participants rated the high explainability AI as significantly less competent ($M = 1.67$, $SE = 0.11$) than they rated the low explainability AI ($M = 1.84$, $SE = 0.11$). Lastly, the interaction effect between AI teammate autonomy and explainability level was non-significant ($F(1, 42) = 0.19$, $p = 0.67$, $\eta^2 < .01$).

These results show that the high autonomy AI teammate was perceived as significantly more competent at completing its task work than the low autonomy AI teammate. These results provide insight into RQ1 by showing participants related less explainability from the AI with a higher level of competence. This result also presents further support for the previous finding on trust. Specifically, it is intriguing that participants felt the less explainable AI was both more competent and more trustworthy than the low explainability AI. This shows that increasing AI explainability is not always appropriate or helpful for teaming. This disconnect between the XAI movement and these results sets up a notable example of how adaptive autonomy may be useful not only in autonomy levels but also in explainability levels, especially when it comes to complex social contexts like teaming.

3.2.3 Perceived awareness of AI teammate actions

The main effect of AI teammate autonomy level on participants' awareness of AI actions was non-significant ($F(1, 42) = 3.15$, $p = 0.08$, $\eta^2 < 0.01$). Furthermore, the main effect of AI teammate explainability level on awareness was also non-significant ($F(1, 42) = 1.99$, $p = 0.17$, $\eta^2 = 0.01$). The interaction effect between AI teammate autonomy and explainability levels on awareness was also non-significant ($F(1, 42) = 0.11$, $p = 0.75$, $\eta^2 < 0.01$) (see Fig. 4).

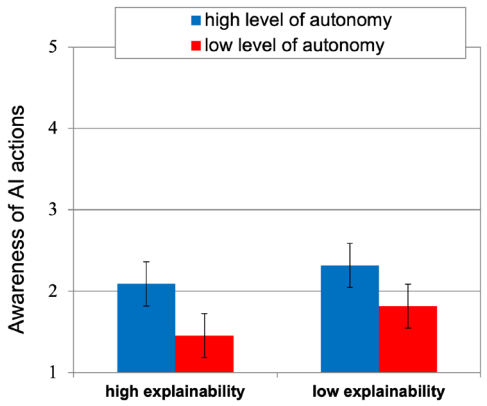

Fig. 4 Awareness of AI actions on a scale of 1 to 5, by the level of autonomy for high and low explainability (error bars represent standard error of the mean)

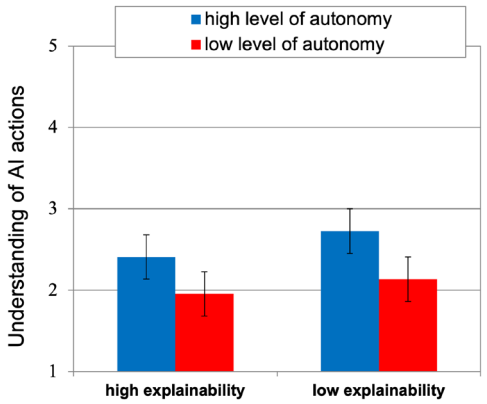

Fig. 5 Understanding of AI actions on a scale of 1 to 5, by the level of autonomy for high and low explainability (error bars represent standard error of the mean)

While participants' awareness of their AI teammate's actions was not significantly affected by the explainability or autonomy level of the AI teammate, values for awareness were higher for participants in the high autonomy condition, a result that should be considered in future work.

3.2.4 Understanding of the AI teammate

The main effect of AI teammate autonomy level on participants' understanding of the AI teammate was non-significant ($F(1, 42) = 2.40, p = 0.13, \eta^2 < 0.01$). Additionally, the main effect of AI teammate explainability level on understanding was non-significant ($F(1, 42) = 1.80, p = 0.19, \eta^2 = .01$). Lastly, the interaction effect between AI teammate autonomy level and explainability level was non-significant ($F(1, 42) = 0.13, p = 0.71, \eta^2 < 0.01$) (see Fig. 5).

These results show that participants in this experiment did not feel that increased explainability or autonomy significantly affected their understanding of the AI teammate's actions. One reason for this may be that the task the participants were asked to perform was familiar to the majority of the participant population, which was targeted for having IT and/or networking experience. It is worth noting that, based on the chat log data, only 14 percent of participants requested any additional explanation from the AIs. This suggests that trust and understanding are not closely tied together when it comes to AI explanations and that increasing one will not directly cause an increase in the other. This discrepancy emphasizes that other factors, such as the content explored in Study 2, are important considerations for designing AI that best supports both human teammate trust and understanding.

4 Study 2

While Study 1 explored how the *amount* of information that an AI teammate provides at different LOAs affects human teammate perceptions (RQ1), it did not address *what* information AI teammates should communicate. The following section details the methods and results of Study 2, which encompassed the two qualitative participatory design sessions and exploration of RQ2.

4.1 Methods

In order to further understand and expand upon the results of Study 1, twelve of the study's participants were recruited to participate in one of two participatory design sessions. Such sessions have been shown to produce realistic, innovative design solutions within the HCI community (Thieme et al. 2023). These participatory design sessions took place over Zoom after the experiment's completion. In this way, participants had a fresh idea of what kinds of teaming scenarios and roles an AI teammate might occupy and the information they would need to provide to human teammates. Study 2 utilized a similar IT networking scenario for the participants in order to explore the content of an AI teammate's explanations.

4.1.1 Participants

All participants who completed Study 1 were asked if they would be willing to participate in a participatory design session relating to the experiment. For our sessions, we decided that a flexible, conversational workshop would most benefit our focus on the needs of the human teammate (Weber et al. 2015). We aimed to schedule five to seven participants per

Table 2 Participatory design session I

Session	Gender	Age	Occupation	Ethnicity
1	Woman	34	Cyber security	American Indian
1	Man	33	Cyber security	Asian
1	Man	29	Network engineering	White
1	Man	33	Software development	White
1	Man	33	Cyber security	White
2	Woman	67	Insurance sales	White
2	Woman	27	Graduate student	White
2	Woman	26	Graduate student	White
2	Man	27	Software development	White
2	Man	69	Electrical engineering (Ret.)	White
2	Woman	66	IT project management (Ret.)	White
2	Man	33	Copyright design	White

session, as we were sensitive to the fact that should the group become too large, it is easy for a few individuals to dominate the conversation (Weber et al. 2015). We provided the initial twenty-three volunteers with the time slots of the sessions, and through this schedule ended up with twelve total participants, five for Session 1 and seven for Session 2. The demographics of these participants are reflected in Table 2.

4.1.2 Design scenario and questions

Prior to the sessions, the lead author conducted a pilot session with three individuals who provided feedback on the details of the scenario and procedure for the session. The primary outcome from the pilot session was the switch to Google Jamboard for collaboration between participants; whereas, in the pilot, the group utilized a shared Google Doc. Both participatory design sessions were conducted over Zoom, with the lead author directing the session. For each session, participants were described the scenario, the design questions, and the schedule for the session. The scenario reflected that of the experiment, in which the participants played the role of IT professionals for an IT help team on a university campus. Their AI teammate was in charge of making software configuration changes to devices connected to the campus network as needed. The participants were told that the agent progressively changed to lower levels of autonomy throughout the incident response cycle, according to a previous study on adaptive AI in incident response (Hauptman et al. 2022). Specifically, the AI teammate would begin the incident response at close to full autonomy and decrease to partial autonomy as the incident response cycle entered the containment phase. Participants were presented with three Design Questions that incorporated what, how, and when AI teammates should explain their decisions to the team:

DQ1: What would you want/need your AI teammate to explain?

DQ2: How would you want/need it communicated (textual, visual, audible, physical methods)?

DQ3: When does the amount of explainability increase/decrease?

4.1.3 Session procedure

Participants received a Zoom invitation and Jamboard link 10 min prior to the session. Once all participants were logged on, the lead author reviewed the consent to the study and received verbal agreement to record the session. After initiating the recording, the lead author reviewed the scenario and design questions and answered any questions from the participants before entering into the semi-structured session. Participants were then asked to individually brainstorm their answers to the design questions (how much and in what way they would want their AI teammate to explain its work and under what conditions that vary) on the Jamboard. Once all the participants announced their completion, the group came together and discussed their thoughts. Throughout this process, the participants changed the notes on the Jamboards, the final products of which were used for analysis. The sessions concluded with the first author reviewing the design questions, the group answers to the questions, and inviting any additional or closing comments.

4.1.4 Qualitative analysis

At the conclusion of the session, the entire Zoom session was automatically transcribed by the Zoom software. The first author then reviewed both recordings by hand to fix transcription errors and provided these copies to the other authors for analysis. The Zoom transcripts and Jamboards were coded using a thematic coding process (Gavin 2008; Braun and Clarke 2012), following which the authors conducted axial coding to develop main themes prevalent in the data (Scott and Medaugh 2017). This reflexive process permitted the study data to guide the analysis (Blair 2015). First, the authors line by line coded the transcripts. Next, these codes were grouped into like categories. Finally, the authors combined groups into large themes that related to the design and research questions. Once these themes were developed, they were considered in concert with the quantitative data from the experiment in order to determine how they did or did not help explain the results of the study, as well as to determine the main themes in how an adaptive AI teammate should functionally explain itself to its teammates. The participatory design sessions were vital to uncovering this portion of the design recommendations, as

they allowed the participants to think through situations and interaction methods beyond what was presented in the experiment. Indeed, as we will show in the following sections, the sessions revealed several important considerations for designing an explainable AI teammate with differing levels of autonomy and explanation.

4.2 Results

The participatory design sessions provided two main artifacts for analysis: the Jamboards, shown in Figs. 6 and Fig. 7, and session transcripts that both help answer the three DQs presented to the participants. DQ1 and DQ2 are nested under the study's RQ1 and DQ3 under RQ2. While the experiment provided some significant data for answering RQ1, it provided no significant data in terms of RQ2. In contrast, by not specifically probing the participants about autonomy level, the participants brought it into their discussions and provided us with ample data that addressed RQ2. From the data collected between the two sessions, we identified seven main themes that the participants agreed upon in regard to the three design questions. In this section, we will use the participants' written and verbal comments to illustrate these themes in detail.

4.2.1 DQ1: Explanations of confidence and situation

Both sessions first focused on the *what* aspect of AI explanations or the main contents that human teammates would need an AI teammate to explain to them. Two main themes emerged from the PD sessions as the most important for an AI teammate to explain: 1) the decision logic behind and the confidence in the AI teammate's decisions; and 2) contents (i.e., terminologies, situation description) that help align teammates' knowledge and understanding of the shared task and situation.

Early in both sessions, participants expressed the desire to see the logic that an AI teammate used to make a decision, believing that it was far more important for the AI to explain the logic path behind its decisions, as opposed to what specific tasks it was doing:

> *"I'm primarily concerned with the specific data points that the AI used to make a determination, as well as the logic that it used" (Male, 27, Software Development).*

> *"I would want a step-by-step indicator of the logic" (Female, 26, Graduate Student).*

As the participants explained above, explanations of the AI's logic path would show the team that the AI possessed the

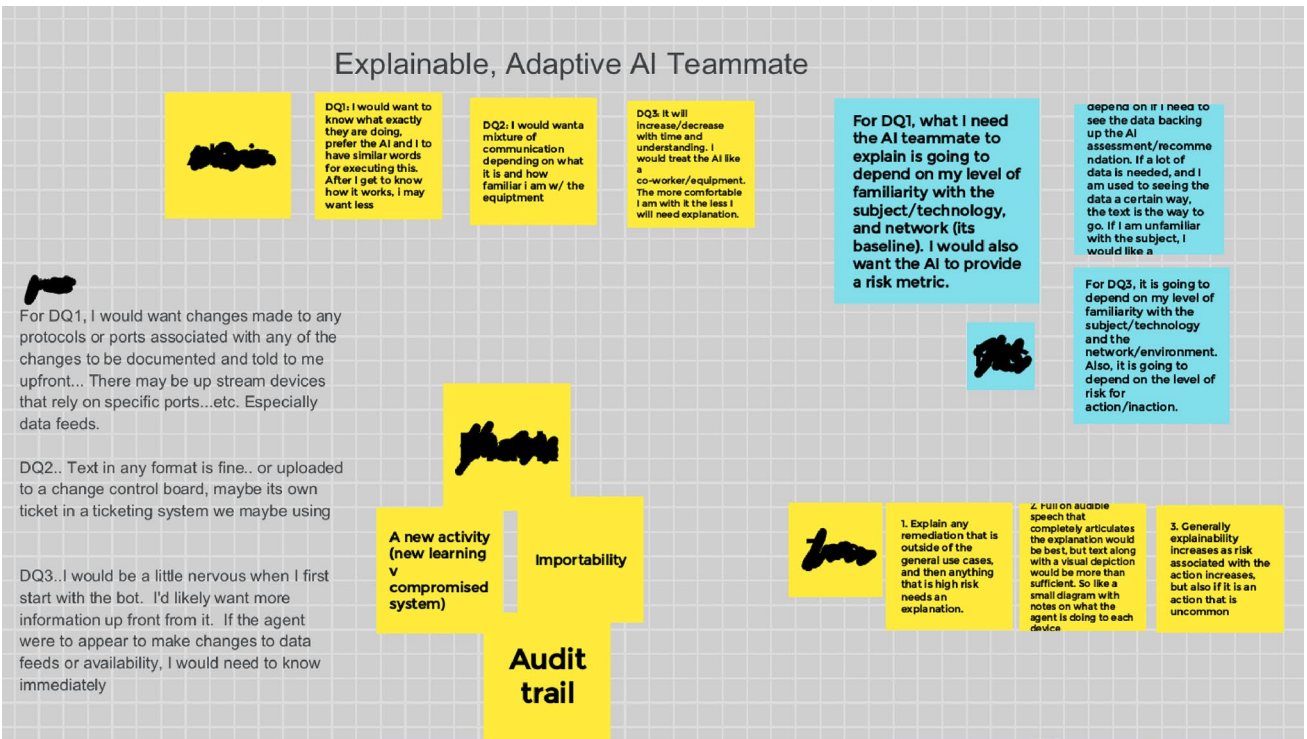

Fig. 6 Participatory design session 1 Jamboard. This session was much more talkative, and the notes were more constructed after the fact. Participants emphasized the importance of building trust and competence in an AI through increased explainability early on in its incorporation into the team

Fig. 7 Participatory design session 2 Jamboard. This group of participants spent a lot of time getting their thoughts on the board before the discussion. This group emphasized the importance of the adaptability of the AI teammate in terms of changing its explainability and autonomy based on risk and team experience

proper data points to be confident in its decision. This element of confidence became an essential topic of discussion that participants reiterated throughout the sessions as the concept of AI autonomy levels came into play. Participants desired for the AI to explain its confidence in its decision through some form of a visual indicator and associate confidence level with the AI's autonomy level. As the participant details in the following quote, she would be more comfortable with an AI operating at a high autonomy level when she can see that its confidence level is high:

> *"In cases where something is very routine, or where it's something new for the AI, going back to the confidence indicator mentioned, it has a set point, some number as configured of confidence, then it can go ahead and do it, and just report the outcomes" (Female, 66, IT Project Management).*

It would help people understand and dynamically adjust whether or not the AI was or should be operating at the appropriate autonomy level if it was able to explain its decisions in terms of confidence levels. For instance, if an AI teammate was operating at a high autonomy level during routine operations and it encountered a new situation in which its confidence level in its decision logic dropped, an explanation of its now lower confidence would communicate to the team that it should be operating at a lower autonomy level because human might want to oversee and intervene in its less confident decisions.

Another important aspect of explainability is that it serves to ensure a shared understanding. For instance, explanations provided to the team should allow team members to align their use and understanding of the meanings of the terminologies:

> *"I really want to make sure our words are similar for execution because words are really important, and I want us to be on the same page" (Female, 34, Cyber Security).*

Indeed, one term could have various meanings depending on the context and disciplinary background. Reversely, individuals from different backgrounds might use different words or terms to mean the same thing. Not sharing the same vocabulary will lead to miscommunication, hindering team effectiveness and efficiency. The need for shared terminology is all the more crucial to generating shared understanding, especially within multi-functional multi-disciplinary teams. Participants wanted the AI to not only explain the terms themselves used to accommodate team members who are not familiar with the terms but also be able to detect and explain those used by human team members. The above

participant used the example of a "computer security survey." While a survey means an examination for computer scientists, psychologists immediately associate it with questionnaires. To quickly align team members' understanding of what actions the AI intends to perform with the survey, the AI needs to account for all the team members' knowledge and background in its explanation.

In addition to aligning team members' understanding of terminology, participants emphasized the importance of aligning team members' awareness of their shared task and situation through AI's explanation:

> *"It provides everyone else with awareness about what's going on so that they can make their own analysis" (Male, 27, Software Development).*

In discussing the needed level of explanation and autonomy, the above participant desired the AI to align everyone's understanding and awareness of their shared situation such that the human teammates can leverage the information and explanation provided by the AI to make their own analysis and judgment to check against that of the AI's. The greater awareness the AI's explanation provides human teammates, the better and the more efficient the latter could give their "informed" input to the AI's decisions and actions.

4.2.2 DQ2: Communication through existing channels and norms

Next, the session conversations turned toward how AI teammates should provide explanations to teammates. These conversations revealed the need for seamless integration of the AI explanations into the team via existing communication platforms and channels (without creating a new one) while its communication style and modality fitting into the team (or organizational) culture:

> *"Whatever kind of thing the rest of the team is currently collaborating on, the ability to seamlessly kind of add that to whatever the AI reports on" (Female, 34, Cyber Security).*

Participants felt that an important aspect of the AI being part of the team, as opposed to just a tool used by the team, is that it communicated in line with the communication methods the team itself uses. The examples of a team Slack channel and Skype were both mentioned in this regard. Similarly, participants explained that the formal explanations of an AI teammate's action should also align with how human teammates document their actions. Within the context of the session scenario, participants suggested that the AI submit its explanations to whatever ticketing system the campus uses for its IT issues:

> *"Whatever recommendation that the agent may have is submitted directly into the ticketing system, so that way as you approve it and stuff, it's logged in the same fashion as everybody else's work" (Male, 33, Cyber Security).*

> *"People are used to seeing certain data to make decisions, and they are used to seeing the data in a certain form" (Male, 33, Copyright Design).*

What both of these participants further emphasized in these quotes and discussion is that mainly when AI teammates are at lower autonomy levels and need humans to make a formal decision based upon their explanations, the recommendations and requests to act need to be submitted in the same format and via the same platform the team uses to consider all of its decisions. In this way, teammates can assess and make good decisions in the same manner they do for their personal tasks.

In addition to AI explanations integrating into team platforms, participants also emphasized the importance of AI explanation methods to fit into team culture:

> *"I feel like the way that the AI is presented with this team should be in conjunction with the way the team interacts with each other" (Female, 27, Graduate Student).*

Participants put this into the perspective of the modality of team interaction. Teams that meet daily in person would respond better to an AI teammate that can communicate through a physical platform, a visual or physical interface. Teams that communicate primarily through online collaboration platforms would respond best to AI teammates with an account and communicate in line with that platform. Most importantly, as this participant so aptly summarizes here, is that whatever modality the AI utilizes to provide its explanations to the team, it needs to be either a collaborative or representative decision:

> *"It has to be a collaborative decision between you and your teammates of like who, what kind of entity would I feel most comfortable with having on my team" (Female, 26, Graduate Student)*

4.2.3 DQ3: One-size explanations do not fit all

The third and final design question that participants considered in the participatory design sessions was under what conditions, if any, they would want the explanation types of the AI teammates to change. In addressing this question, participants were almost all opposed to the idea that an AI teammate's amount of explanation and its autonomy level

follow a simple linear relationship; instead, the amount and the timing of AI teammate explanations should be based on the human teammates' moment-by-moment need to know. As the following participant emphasized, the AI needs to be able to make certain assumptions about what its teammates already know in order not to overload them with excessive and redundant explanations.

> *"I think that regardless of what the agent is communicating, I think it needs to be very efficient responses, not too wordy because that could be a lot as well, so I feel like the agent has to assume the user has some level of expertise" (Female, 27, Graduate Student)*

There's a trade-off between explaining too much and too little. On the one hand, the AI should be as brief as possible so as to not annoy everyone by explaining every single thing it does. Conversely, the AI needs to provide enough information and explanation so everyone on the team can make proper sense of it. It requires the AI to make accurate assumptions about what its human teammates already know to provide an appropriate level of explanation. While humans can make relatively correct assumptions about what other humans know (Fussell and Krauss 1992), it is yet to be configured into the AI to possess such ability.

Additionally, the amount a person knows, and thus the amount of explanation needed, depends on various factors such as their disciplinary background and experience. The AI certainly would not need to explain "NLP" to a computer scientist but would need to do so for lay people or someone newly onboard. Furthermore, as team members aggregate information during their interaction and collaboration, the need to explain previously mentioned ideas should decrease. Another participant echoed this, stating that repeated over-explanations by the AI can quickly lead to human frustration:

> *"You don't want to have a situation where it asks, and I've told you once now, and I gave you new ground, I told you a second time, and if you ask the same stupid question the third time, you know, I'm going to be pissed" (Male, 69, Electrical Engineering).*

Another aspect that affected the need for humans to know was the risk and complexity of the AI's task. Participants indicated that the higher risk and/or more complex an action an AI teammate would take is, the more they would need the AI to explain its decision to the rest of the team.

> *"Explainability increases as both risk and complexity of the action increases" (Male, 29, Network Engineering).*

Similar to the idea of the confidence indicator, a value should be assigned to the risk associated with a task. With higher levels of risk, the AI needs to explain more to the team. This aligns with teamwork research that shows the need for increased communication between team members as task risk increases (Leonard et al. 2004).

As participants considered the third design question, they were again asked to consider how AI autonomy level played into when and how much explanation is needed, if at all. The collective opinions were that in terms of autonomy level, textual and written explanations are required to increase as an AI's autonomy level increases. The main reason for this is for auditing purposes:

> *So, you can very easily reconstruct whatever issue happened from there, so it makes the entire post-analysis process a lot easier" (Female, 67, Insurance Sales).*

Participants discussed the reality that both people and AI make mistakes or actions result in unintended consequences, so there may be the need to go back and understand an action that an AI took, mainly when it operated at a heightened autonomy level with less human input into the decision. Ultimately, a human team leader will always be responsible for the actions an AI takes under their lead. For this reason, as an AI teammate operates with less and less human oversight and input, it is even more important that the explanations it makes of its actions be made in an auditable, textual manner that its teammates can consult both during and after it takes action.

5 Discussion

In this study, we sought to explore how changes in an AI teammate's level of explainability and level of autonomy individually and jointly affect a human teammate's trust in and perception of the AI. To do this, we conducted a WoZ experiment in which the participant worked with an AI teammate with varying levels of explainability to complete IT networking tasks, after which we invited some of the participants to take part in participatory design sessions to further clarify the design implications of the experiment. The discussion will address that, in regards to RQ1 concerning how human teammate perceptions change as AI explainability increases, in some teaming situations increasing explainability is actually counterproductive to creating a trustworthy, competent AI teammate. This is because human teammates want AI communications only at appropriate intervals, such that they don't interfere with their own tasks. It will also address, in regards to RQ2, how these perceptions change as an AI LOA changes, that LOA itself does not affect a teammate's explainability needs. In fact, the explainability needs of a human teammate may actually contribute to selecting the optimal LOAs for the AI teammate.

5.1 Explainability and the need to know

The experiment and the participatory design studies presented in this paper jointly provide interesting insights into how explainability and autonomy levels dynamically influence human perceptions of their AI teammates and suggest important considerations for the information science community. The 2x2 experiment produced counter-intuitive results (e.g., the high explainability agent was perceived as less competent and less trustworthy) that required additional insight in order to understand why lower explainability AI teammates were perceived as more trustworthy and competent. The participatory design sessions revealed four main factors that influence the team's "need to know," which determines the level of explainability needed from an AI teammate. These four factors are 1) the AI's confidence in its decision logic, 2) the absolute task complexity and task complexity relative to human familiarity, 3) risk, and 4) the ability to audit the AI's explanations.

The idea of a confidence indicator was prevalent in the participatory design sessions, most evident in the session 1 Jamboard shown in Fig. 6. The participants emphasized that an indicator of the AI's confidence in its decision logic can provide them with an easy and efficient heuristic to trust that it's doing the right thing. This makes sense given the reasons why AI teammates are attractive: they can handle large sets of data and computational workloads beyond human capability (Duan et al. 2019). Because the AI should possess superior processing capabilities, it is logical that a human teammate would prefer to be told how confident the AI teammate is in its decision than having to try to make sense of its explanations of those heavy computations, a language barrier that drives other communication needs such as natural language processing (Zhuang et al. 2017). This indicator would show a human teammate how much input the AI needs from them, and mirrors recent HCI research that shows humans desire more explanations that help them collaborate better, as opposed to just what the AI is doing (Kim et al. 2023).

The second aspect that participants identified as affecting their need to know was task complexity relative to a human teammate's familiarity and experience with the task. Research in all-human teams has shown that task complexity heavily influences individual perceptions of teammates and intra-team trust (Choi and Cho 2019), and recent studies into human-robot teams have shown similar degradation of trust as the task complexity increases (Krausman et al. 2022; Zhang et al. 2023). Our study supports a relationship between task complexity and the perceptions of teammates in human-AI teams. The fact that the task complexity in our experiment was not very high made our participants feel that the AI teammate with lower explainability was more competent because a task as simple as this does not require much explanation. Additionally, this notion of task complexity relative to human teammate's familiarity and experience with the task differs from the absolute task complexity, as it must account for an individual's knowledge and experience with the task that might differ across teammates and can change moment-by-moment. This aligns with current HCI research showing that the role AI plays in supporting a human should consider user expertise and task complexity, and intelligent systems need to be capable of discriminating between different users (Buckland and Florian 1991). As our participants pointed out, a thorough explanation is desirable only for the first time it is needed; it becomes annoying and even detrimental to the team dynamic if the AI repeatedly explains the same thing regardless of whether it is needed or not. However, newer teammates or newer tasks require more explanation from the AI teammate, as the human teammates are still trying to understand what and why the AI is doing something under these new conditions and, consequently, how its actions affect the actions they themselves are taking. This is in line with explainable AI research into question-based explainability, which has shown users of different experience levels will have differing explainability needs to be based upon questions they are likely to ask (or not ask) (Liao et al. 2020). Therefore, the AI teammate's explainability should be tailored to the human teammates' moment-by-moment needs. The more complex the task is to human teammates (due to lack of prior knowledge or experience), the greater the need to know.

Third, the risk level associated with the task is just as important. Participants discussed the risk of the AI agent's actions as a defining factor in the team's need to know because as the risk increases, the actions that an AI teammate is taking are more likely to affect the actions of the team at large. High-risk actions may bear additional considerations, such as the ethical concern of having too little human reason involved in the decision process (Shneiderman 2020; Tolmeijer et al. 2022). Indeed, risk decision literature shows that in evaluating decisions made in uncertain conditions, two of the most vital questions to ask are 1) what are the potential impacts of that decision, and 2) is the decision ethically good (Ersdal and Aven 2008)? Our session participants discussed this in terms of how you would expect human teammates to pause and take extra care to explain to the team and their leaders what and why they intend to do something in a high-risk situation. The riskier the decision, the more explanation and consideration it requires.

The fourth factor that affects the team's need to know is the ability to return to and audit an AI agent's explanations. In discussing how teams would need to receive explanations from an AI teammate at different autonomy levels, the participants repeatedly returned to the idea of integrating the AI's explanations into the team's existing communication and auditing platforms, particularly in design session two, shown in Fig. 7. In the sessions, there

was often a tug of war between participants in not wanting to be distracted by their AI teammates and not trusting them to always make the right decision. This clash was fueled by the reality that if an AI teammate makes a wrong decision or the team performs poorly, then at the end of the day, it is the human teammates who will be held responsible. This concept of human accountability is widely considered the first principle of AI ethics (Lim and Kwon 2021). Thus, there is a requirement for human teammates to understand what and why an AI teammate made certain decisions not only before but also during and after the fact. Participants explained that the easier these explanations are to review and audit, the more comfortable they would be with a more autonomous teammate. This concept can be considered an extension of the auditing processes placed on human employees/teammates to decrease insider threats (Colwill 2009). Organizations require humans to document and explain their actions in a recordable format, such that if there is a question of their trustworthiness or recklessness, teammates and superiors can review those explanations of actions. Likewise, it would be prudent to enforce similar auditing mechanisms for autonomous teammates. This reinforces previous HCI research that has established the need for traceability of an AI agent's decisions in order to ensure there is proper accountability for any repercussions of its actions (Lim and Kwon 2021).

This desire to receive communications only at certain times helps explain our experiment results, where our participants actually rated the low explainability agent as more trustworthy and competent. As the AI provides more explanations to a human teammate during a team task, it appears to want more human input into its actions. Thus, even though participants would like to be able to review the AI teammate's actions, its constant communications decrease its apparent independence. The factors identified above may be key to helping to resolve this issue, as they identify the most important times and types of explanations that AI teammates should provide. Targeted, adaptive explanations promote trust in the AI by providing reasoning to human teammates when they need it most, without overloading or annoying them with information unrelated to their own team role. Combined, these four factors represent the team's need to know an explanation from an AI teammate in media res (in the middle of things). This need to know can also be used to inform the optimal levels of autonomy for an AI teammate, as we will describe in the following design recommendations.

5.2 Design recommendations

We will now present two important design recommendations that should be incorporated into the future development of AI teammates. The first of these recommendations focuses on the concept of adaptive explainability based on a team's need to know. The second recommendation is that AI teammates can operate at higher LOAs when certain explainability conditions are met.

5.2.1 AI teammates should exhibit adaptive explainability based on their team's need to know

The first recommendation derived from the results of these studies is that AI teammates need to assess a task's complexity and risk level and use these values to determine the team's need to know. This need to know, as discussed in the previous section, determines the level of explainability the AI teammate provides over time. In terms of the scenario utilized in the participatory design sessions, the AI teammate could read an incident ticket submitted to the team and determine how complex its response actions would need to be and the risk level of negative impacts to the campus network of those actions. Using a predefined decision matrix to aggregate the values (as suggested by one of our participants), the AI teammate, going into the incident, would know the appropriate level of explainability to provide to its teammates in responding to this specific incident ticket. Additionally, the AI's user interface should display a visual indicator of this aggregated complexity-risk value. This informs the rest of the team what level of explainability to expect from their teammate over the course of the task. For instance, if the way the team interacts with the autonomous teammate is over a team chat channel, there could be an explainability icon next to its avatar indicating this value. This recommendation has several implications for the HCI community, particularly in regard to AI interface design. This recommendation specifically charges AI designers to create interfaces that allow for side-by-side displays of AI explanations with this complexity-risk value.

5.2.2 AI teammates should have higher autonomy when they can provide high levels of written, archival explainability

When AI teammates possess lower levels of autonomy, human teammates have more time to consider and process the explanations, ask for more details if necessary, and take notes of the explanations as needed. In other words, there is ample time for humans to understand why their AI teammate is doing something and how that is going to affect the team. When AI teammates operate with higher levels of autonomy, the time to understand the AI's decision is shortened, and teammates may need to go back and consult those explanations during or after a task. These explanations are key to providing a degree of accountability over the AI (Raji et al. 2020). For this reason, written, archival

explanations can allow AI teammates to operate at higher levels of autonomy because their explanations are available to the team for an extended period of time. An example of this within this study's context would be if the AI provides explanations to the team through an online ticketing system. When the software agent is connected to a network and able to submit these explanations to the system as it conducts its actions, it could possess higher levels of autonomy because the rest of the team is able to consult the explanations in the system at any time. If the autonomous teammate is deployed on a computer system that is not currently connected to the network and can only communicate through a chat resident on that volatile system, it would need to operate at a lower level of autonomy, as the longevity of its explanations likely depends on the awareness and note-taking of one of its teammates. In this way, the ability of the autonomous teammate to provide archival explanations serves as a guiding factor in determining its optimal autonomy level(s). For the HCI community, this is an important design recommendation that requires organizations to consider not only the tasks that an AI teammate will perform but also when and where it will perform them. AI utilized for more than one team or operating environment may need to be capable of communicating on several platforms, such that it can be adjusted to use whichever method is preferred by the team at the time. It would then also need to be able to be deployed at multiple levels of autonomy, based upon the types of explanations selected.

5.3 Limitations and future work

The research presented in this paper has limitations that should be considered when interpreting these results. First, the autonomous teammate with whom our participants interacted during the experiment communicated purely over text. The cognitive effort to process the textual explanations may be greater than that to hear the explanations using auditory methods. Future research should further explore the effects of different forms of explanations, such as auditory and pictorial, as research shows that there are significant advantages to using explanation methods that cater more specifically to individual learning needs (Wolf and Ringland 2020). Second, our studies involved only one autonomous teammate, and it will be important to assess how different team compositions of more autonomous and human teammates affect how humans perceive the explanations of their autonomous teammates, as human-AI teaming research has already shown that team composition affects a variety of teaming factors (Schelble et al. 2022a). For instance, while consulting multiple text explanations from a single AI teammate may be helpful to the team, three providing real-time textual explanations may overwhelm the team's cognitive capabilities while focusing on their own tasks, a further factor to consider in equipping agents with adaptive explainability. In terms of our participants, it is worth noting that our participatory design sample contained only two ethnic minorities, thus presenting a largely white perspective on the questions. A more diverse pool may present additional important findings. Finally, the experiment and design sessions in this paper focused on a singular, relatively low-risk task and environment. Because the results of this study indicate that task complexity and risk have a lot of bearing on optimal explainability levels, it will be important that the community studies the effects of varying task and risk complexity.

6 Conclusion

As humans and autonomous agents collaborate and work more independently as teammates, the explanations that autonomous agents provide to their human counterparts become more and more critical. In this study, we showed that the level and type of explainability an artificially intelligent agent provides significantly affect the team's perceived competence and trust in that agent. Counter-intuitively, the participants in our teaming scenario perceived the AI agent with a lower level of explainability as more trustworthy and competent than one with a high level of explainability. Our participatory design sessions helped explore this paradox and guided our creation of two crucial design recommendations for the HCI community concerning the details, frequency, and modality of AI explanations centered on the unique concept of adaptive explainability. These recommendations enable the information science community to model and design adaptable AI agents that humans can perceive as capable, trustworthy teammates at varying levels of autonomy.

Appendix A

Data analysis of the significant effects of AI explainability level (Figs. 8, 9, 10, 11, 12 and 13).

Fig. 8 F statistics for trust model

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	42	201.573	<.001
LOA	1	42	.299	.587
explanation_level	1	42	4.421	.042
LOA * explanation_level	1	42	.573	.453

a. Dependent Variable: F_trust.

Fig. 9 Pairwise comparisons of explainability levels on trust in the AI

Pairwise Comparisons[a]

(I) explanation_level	(J) explanation_level	Mean Difference (I-J)	Std. Error	df	Sig.[c]
low	high	.189*	.090	42	.042
high	low	-.189*	.090	42	.042

Fig. 10 Effect of explainability levels on trust in the AI

Estimates[a]

explanation_level	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
low	1.962	.139	51.716	1.683	2.241
high	1.773	.139	51.716	1.494	2.052

a. Dependent Variable: F_trust.

Fig. 11 F Statistics for competence model

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	42.000	278.984	<.001
LOA	1	42.000	1.498	.228
explanation_level	1	42.000	5.729	.021
LOA * explanation_level	1	42.000	.189	.666

a. Dependent Variable: F_competence.

Fig. 12 Pairwise comparisons of explainability levels on perceived competency of the AI

Pairwise Comparisons[a]

(I) explanation_level	(J) explanation_level	Mean Difference (I-J)	Std. Error	df	Sig.[c]
low	high	.167*	.070	42.000	.021
high	low	-.167*	.070	42.000	.021

Fig. 13 Effect of explainability levels on perceived competency of the AI

Estimates[a]

explanation_level	Mean	Std. Error	df	95% Confidence Interval	
				Lower Bound	Upper Bound
low	1.841	.111	51.086	1.618	2.063
high	1.674	.111	51.086	1.452	1.897

a. Dependent Variable: F_competence.

Funding Open access funding provided by the Carolinas Consortium. Open access funding provided by the Carolinas Consortium.

Data availability Anonymized data can be made available upon request by contacting the first author.

References

Abbass HA (2019) Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. Cogn Comput 11(2):159–171

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Inform Fusion 58:82–115

Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, and Weld D (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, 1–16

Blair E (2015) A reflexive exploration of two qualitative data coding techniques. J Methods Meas Soc Sci 6(1):14–29

Braun V, Clarke V (2012). Thematic analysis

Buckland MK, Florian D (1991) Expertise, task complexity, and artificial intelligence: A conceptual framework. J Am Soc Inform Sci 42(9):635–643

Caldwell S, Sweetser P, O'Donnell N, Knight MJ, Aitchison M, Gedeon T, Johnson D, Brereton M, Gallagher M, Conroy D (2022) An agile new research framework for hybrid human-ai teaming: Trust, transparency, and transferability. ACM Trans Inter Intell Syst 12(3):1–36

Castelvecchi D (2016) Can we open the black box of ai? Nature News 538(7623):20

Centeio Jorge C, Tielman ML, Jonker CM (2022) Artificial trust as a tool in human-ai teams. In Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, 1155–1157

Chen Z (2023) Collaboration among recruiters and artificial intelligence: removing human prejudices in employment. Cogn Technol Work 25(1):135–149

Chen J, Sun J, Wang G (2022) From unmanned systems to autonomous intelligent systems. Engineering 12:16–19

Chignell M, Wang L, Zare A, Li J (2023) The evolution of hci and human factors: Integrating human and artificial intelligence. ACM Trans Comp Human Inter 30(2):1–30

Choi O-K, Cho E (2019) The mechanism of trust affecting collaboration in virtual teams and the moderating roles of the culture of autonomy and task complexity. Comp Human Behav 91:305–315

Cohen SN, Snow D, Szpruch L (2021) Black-box model risk in finance. arXiv preprint arXiv:2102.04757

Cohen J (1988) Statistical power analysis for the behavioral sciences. Academic press, Newyork

Colwill C (2009) Human factors in information security: The insider threat-who can you trust these days? Inform Secur Tech Report 14(4):186–196

Costa AC, Fulmer CA, Anderson NR (2018) Trust in work teams: An integrative review, multilevel model, and future directions. J Organ Behav 39(2):169–184

Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371

Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021) Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif Intell 299:103525

de Lemos R, Grześ M (2019) Self-adaptive artificial intelligence. In 2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS), IEEE, 155–156

Dhanorkar S, Wolf CT, Qian K, Xu A, Popa L, Li Y (2021) Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle. In Designing Interactive Systems Conference 2021:1591–1602

Duan Y, Edwards JS, Dwivedi YK (2019) Artificial intelligence for decision making in the era of big data-evolution, challenges and research agenda. Int J Inf Technol 48:63–71

Ehsan U, Riedl M (2019) On design and evaluation of human-centered explainable ai systems. Glasgow'19

Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. Human Factors 37(1):32–64

Ersdal G, Aven T (2008) Risk informed decision-making and its ethical basis. Reliab Eng Syst Saf 93(2):197–205

Fussell SR, Krauss RM (1992) Coordination of knowledge in communication: effects of speakers' assumptions about what others know. J Pers Soc Psychol 62(3):378

Gavin H (2008) Thematic analysis. Understanding research methods and statistics in psychology, 273–282

Gieselmann M, Sassenberg K (2023) The more competent, the better? the effects of perceived competencies on disclosure towards conversational artificial intelligence. Social Sci Comp Rev 41(6):2342–2363
Ha T, Kim S, Seo D, Lee S (2020) Effects of explanation types and perceived risk on trust in autonomous vehicles. Trans Res Part F 73:271–280
Hauptman AI, Schelble BG, McNeese NJ, Madathil KC (2022) Adapt and overcome: Perceptions of adaptive autonomous agents for human-ai teaming. Computers in Human Behavior, 107451
Hussain F, Hussain R, Hossain E (2021) Explainable artificial intelligence (xai): An engineering perspective. arXiv preprint arXiv:2101.03613
Huvila I, Enwald H, Eriksson-Backa K, Liu Y-H, Hirvonen N (2022) Information behavior and practices research informing information systems design. J Assoc Inf Sci Technol 73(7):1043–1057
Jacovi A, Marasović A, Miller T, Goldberg Y (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, 624–635
Janitor J, Jakab F, Kniewald K (2010) Visual learning tools for teaching/learning computer networks: Cisco networking academy and packet tracer. In 2010 Sixth international conference on networking and services, IEEE, 351–355
Jarrahi MH, Lutz C, Boyd K, Oesterlund C, Willis M (2022). Artificial intelligence in the work context
Joyce DW, Kormilitzin A, Smith KA, Cipriani A (2023) Explainable artificial intelligence for mental health through transparency and interpretability for understandability. Digital Med 6(1):6
Kelley JF (2018) Wizard of oz (woz) a yellow brick journey. J Usability Stud 13(3):119–124
Kim SS, Watkins EA, Russakovsky O, Fong R, Monroy-Hernández A (2023) " help me help the ai": Understanding how explainability can support human-ai interaction. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17
Kosch T, Welsch R, Chuang L, Schmidt A (2023) The placebo effect of artificial intelligence in human-computer interaction. ACM Trans Comp Human Inter 29(6):1–32
Krausman A, Neubauer C, Forster D, Lakhmani S, Baker AL, Fitzhugh SM, Gremillion G, Wright JL, Metcalfe JS, Schaefer KE (2022) Trust measurement in human-autonomy teams: Development of a conceptual toolkit. ACM Transactions on Human-Robot Interaction
Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What do we want from explainable artificial intelligence (xai)?-a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. Artif Intell 296:103473
Larsson S, Heintz F (2020) Transparency in artificial intelligence. Internet Policy Rev 9(2):10
Leonard M, Graham S, Bonacum D (2004) The human factor: the critical importance of effective teamwork and communication in providing safe care. BMJ Quality Safety 13(suppl 1):i85–i90
Liao QV, Gruen D, Miller S (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, 1–15
Lim JH, Kwon HY (2021) A study on the modeling of major factors for the principles of ai ethics. In DG. O2021: The 22nd Annual International Conference on Digital Government Research, 208–218
Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57
Liu J, Marriott K, Dwyer T, Tack G (2023) Increasing user trust in optimisation through feedback and interaction. ACM Trans Comp Human Inter 29(5):1–34
Lombrozo T (2006) The structure and function of explanations. Trends Cogn Sci 10(10):464–470
Lumineau F (2017) How contracts influence trust and distrust. J Manage 43(5):1553–1577
Marcinkevičs R, Vogt JE (2020) Interpretability and explainability: A machine learning zoo mini-tour. arXiv preprint arXiv:2012.01805
McGee ET, McGregor JD (2016) Using dynamic adaptive systems in safety-critical domains. In Proceedings of the 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, 115–121
McNeese NJ, Demir M, Cooke NJ, Myers C (2018) Teaming with a synthetic teammate: Insights into human-autonomy teaming. Human Factors 60(2):262–273
Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38
Mueller ST, Hoffman RR, Clancey W, Emrey A, Klein G (2019) Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. arXiv preprint arXiv:1902.01876
Nakahashi R, Yamada S (2021) Balancing performance and human autonomy with implicit guidance agent. Front Artif Intell 4:142
Nyre-Yu, M., Gutzwiller, R. S., and Caldwell, B. S. (2019). Observing cyber security incident response: qualitative themes from field research. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 63, pages 437–441. SAGE Publications Sage CA: Los Angeles, CA
O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (2020). Human–autonomy teaming: A review and analysis of the empirical literature. Human Factors, page 0018720820960865
Paleja R, Ghuy M, Ranawaka Arachchige N, Jensen R, Gombolay M (2021) The utility of explainable ai in ad hoc human-machine teaming. Adv Neural Inform Process Syst 34:610–623
Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern Part A 30(3):286–297
Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern Part A 30(3):286–297
Pedreschi D, Giannotti F, Guidotti R, Monreale A, Ruggieri S, Turini F (2019) Meaningful explanations of black box ai decision systems. In Proceedings of the AAAI conference on artificial intelligence 33:9780–9784
Pokam R, Debernard S, Chauvin C, Langlois S (2019) Principles of transparency for autonomous vehicles: first results of an experiment with an augmented reality human-machine interface. Cogn Technol Work 21:643–656
Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency, 33–44
Richards D, Stedmon A (2017) Designing for human-agent collectives: display considerations. Cogn Technol Work 19:251–261
Sanneman L, Shah JA (2020) A situation awareness-based framework for design and evaluation of explainable ai. In International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer, 94–110
Schelble BG, Flathmann C, McNeese NJ, Freeman G, Mallick R (2022a) Let's think together! assessing shared mental models, performance, and trust in human-agent teams. Proceedings of the ACM on Human-Computer Interaction, 6(GROUP):1–29

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., and Freeman, G. (2022b). Towards ethical ai: Empirically investigating dimensions of ai ethics, trust repair, and performance in human-ai teaming. Human Factors, page 00187208221116952

Schoenherr JR, Abbas R, Michael K, Rivas P, Anderson TD (2023) Designing ai using a human-centered approach: Explainability and accuracy toward trustworthiness. IEEE Trans Technol Soc 4(1):9–23

Scott C, Medaugh M (2017) Axial coding. The international encyclopedia of communication research methods 10:9781118901731

Shin D (2020) User perceptions of algorithmic decisions in the personalized ai system: Perceptual evaluation of fairness, accountability, transparency, and explainability. J Broadcasting Electron Media 64(4):541–565

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. Int J Human Comp Stud 146:102551

Shin D (2021) Why does explainability matter in news analytic systems? proposing explainable analytic journalism. J Stud 22(8):1047–1065

Shin D (2022) The perception of humanness in conversational journalism: An algorithmic information-processing perspective. New Media Soc 24(12):2680–2704

Shneiderman B (2020) Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. ACM Trans Interactive Intell Syst 10(4):1–31

Slota SC, Fleischmann KR, Greenberg S, Verma N, Cummings B, Li L, Shenefiel C (2022) Locating the work of artificial intelligence ethics. J Assoc Inform Sci Technol 74:311–322

Sokol K, Flach P (2020) Explainability fact sheets. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM

Speith T (2022) A review of taxonomies of explainable artificial intelligence (xai) methods. In 2022 ACM Conference on Fairness, Accountability, and Transparency, 2239–2250

Stowers K, Brady LL, MacLellan C, Wohleber R, Salas E (2021) Improving teamwork competencies in human-machine teams: Perspectives from team science, Front Psychol 1669

Suzanne Barber K, Goel A, Martin CE (2000) Dynamic adaptive autonomy in multi-agent systems. J ExperimenTheoretical Artif Intell 12(2):129–147

Thieme A, Hanratty M, Lyons M, Palacios J, Marques RF, Morrison C, Doherty G (2023) Designing human-centered ai for mental health: Developing clinically relevant applications for online cbt treatment. ACM Trans Comp Human Inter 30(2):1–50

Tolmeijer S, Christen M, Kandul S, Kneer M, Bernstein A (2022) Capable but amoral? comparing ai and human expert collaboration in ethical decision making. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–17

Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. arXiv preprint arXiv:2006.00093

von Eschenbach WJ (2021) Transparency and the black box problem: Why we do not trust ai. Philos Technol 34(4):1607–1622

Waltl B, Vogl R (2018) Increasing transparency in algorithmic-decision-making with explainable ai. Datenschutz und Datensicherheit-DuD 42(10):613–617

Wang N, Pynadath DV, Hill SG (2016) The impact of pomdp-generated explanations on trust and performance in human-robot teams. In Proceedings of the 2016 international conference on autonomous agents & multiagent systems, 997–1005

Wang D, Yang Q, Abdul A, Lim BY (2019) Designing theory-driven user-centric explainable ai. In Proceedings of the 2019 CHI conference on human factors in computing systems, 1–15

Weber S, Harbach M, Smith M (2015) Participatory design for security-related user interfaces. Proc, USEC, 15 pp

Weitz K, Schiller D, Schlagowski R, Huber T, André E (2019) " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 7–9

Wickens CD, Li H, Santamaria A, Sebok A, Sarter NB (2010) Stages and levels of automation: An integrated meta-analysis. In Proceedings of the human factors and ergonomics society annual meeting, vol. 54, Sage Publications Sage CA: Los Angeles, CA, 389–393

Wilson HJ, Daugherty PR (2018) Collaborative intelligence: Humans and ai are joining forces. Harvard Business Review 96(4):114–123

Wolf CT, Ringland KE (2020) Designing accessible, explainable ai (xai) experiences. ACM SIGACCESS Access Comput 125:1–1

Xie SL, Gao Y, Han R (2022) Information resilient society in an ai world-is xai sufficient? Proc Assoc Inform Sci Technol 59(1):522–526

Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J (2019) Explainable ai: A brief survey on history, research areas, approaches and challenges. In CCF international conference on natural language processing and Chinese computing, Springer, 563–574

Yu R, Alì GS (2019) What's inside the black box? ai challenges for lawyers and researchers. Legal Inform Manage 19(1):2–13

Zhang Y, Li Z, Guo H, Wang L, Chen Q, Jiang W, Fan M, Zhou G, Gong J (2023) " i am the follower, also the boss": Exploring different levels of autonomy and machine forms of guiding robots for the visually impaired. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–22

Zhou J, Chen F (2019) Towards trustworthy human-ai teaming under uncertainty. In IJCAI 2019 workshop on explainable AI (XAI)

Zhuang Y-T, Wu F, Chen C, Pan Y-H (2017) Challenges and opportunities: from big data to knowledge in ai 2.0. Front Inform Technol Electron Eng 18(1):3–14

Zieba S, Polet P, Vanderhaegen F, Debernard S (2010) Principles of adjustable autonomy: a framework for resilient human-machine cooperation. Cogn Technol Work 12(3):193–203