Cognition, Technology & Work
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

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

Published online: 18 May 2024

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