

Relationship Between Trust in AI Creator and Trust in AI Systems: The Crucial Role of AI Alignment and Steerability

Kambiz Saffarizadeh*

Department of Information Systems and Operations Management, College of Business,
University of Texas at Arlington, 701 S W Steet, Arlington, TX 76010, USA
e-mail: kambiz.saffari@uta.edu

Mark Keil

Department of Computer Information Systems, J. Mack Robinson College of Business, Georgia
State University, 55 Park Place, Atlanta, GA 30303, USA
e-mail: mkeil@gsu.edu

Likoebe Maruping

Department of Computer Information Systems, J. Mack Robinson College of Business, Georgia
State University, 55 Park Place, Atlanta, GA 30303, USA
e-mail: lmaruping@gsu.edu

The Version of Record of this manuscript has been published and is available in the Journal of Management
Information Systems (2024) <https://www.tandfonline.com/doi/10.1080/07421222.2024.2376382>

ABSTRACT

This paper offers a novel perspective on trust in artificial intelligence (AI) systems, focusing on the transfer of user trust in AI creators to trust in AI systems. Using the agentic IS framework, we investigate the role of AI alignment and steerability in trust transference. Through four randomized experiments, we probe three key alignment-related attributes of AI systems: creator-based steerability, user-based steerability, and autonomy. Results indicate that creator-based steerability amplifies trust transference from AI creator to AI system, while user-based steerability and autonomy diminish it. Our findings suggest that AI alignment efforts should consider the entity with which the AI goals and values should be aligned and highlight the need for research to theorize from a triadic view encompassing the user, the AI system, and its creator. Given the diversity in individual goals and values, we recommend that developers move beyond the prevailing 'one-size-fits-all' alignment strategy. Our findings contribute to trust transference theory by highlighting the boundary conditions under which trust transference breaks down or holds in the emerging human-AI environment.

Keywords: The AI Alignment Problem, Trust Transference, Creator-Based Steerability, User-Based Steerability, Autonomy, Algorithmic Decision Making, AI Ethics

INTRODUCTION

As companies adopt artificial intelligence (AI) for various business tasks [58], from screening résumés [13] to stock market trading [14] and medical diagnosis [43,49], the role of trust in AI systems in general and agents in particular becomes paramount. However, concerns have arisen about the trustworthiness of the technology companies creating these agents (hereafter referred to as AI creators) [10]. With reports of misuse of personal information and low trust in technology companies among the public [26], it is crucial for managers to consider the potential erosion of users' trust in the AI creator and how this might affect trust in the creator's AI agent. This matters because the efficacy, cost-cutting, and revenue-generating potential of AI agents hinges on the degree to which intended users of these agents trust them in their daily decision-making tasks [24].

Trust transference theory suggests that trust in an AI creator can be transferred to its AI agent when there is a perceived meaningful association between them [89]. However, as AI agents become more advanced, AI alignment becomes more challenging [30] and the association between an AI creator and AI agent becomes less clear. AI alignment is the process of ensuring that AI agents act in accordance with human goals and values [30,31].¹ For instance, Microsoft had to apologize for racist and misogynistic tweets posted by its autonomous AI agent and argued that the agent's tweets did not represent Microsoft's values [62]. Such scenarios underscore the importance of AI alignment in understanding people's trust in AI agents considering their trust in AI creators.

The importance of AI alignment with human goals and values and the impact it may have on trust in AI agents has been recognized by practitioners [29,30] and researchers in various areas of inquiry such as machine morality [8,56], AI ethics [19], interpretability [11], explainability [77], and transparency [101]. However, this extant work has not addressed the possibility that relevant

¹ While AI alignment may seem similar to the concept of person-organization value congruence [39] in the management literature, these concepts are different in that they deal with different types of relationships (e.g., leader-employee versus user-AI agent) and different entities (e.g., humans versus humans and non-humans).

stakeholders may hold different goals and values [31], raising the question of whose goals and values the AI agent should align with. This issue is critical as various stakeholders, such as AI creators and users, can influence the AI agent to align with their specific goals and values. When these stakeholders have conflicting value systems, such multifaceted alignment can adversely affect trust in the AI agent. For instance, after interacting with ChatGPT, a GPT-based chatbot created by OpenAI, users with conservative political views expressed their belief that ChatGPT was embedded with OpenAI's biases against conservative views (a perceived conflict between the AI creator's values and users' own values) [64]. This example highlights the complexity of AI alignment as people may have different views of the same problem based on their value system [4]. Sam Altman, the CEO of OpenAI, drew attention to this issue as one of the most pressing challenges they faced in the development of their AI agents [93].

To unpack the AI alignment problem, we draw from the agentic IS framework [2], which yields a triadic view of human-AI interaction involving the AI creator, the AI user, and the AI agent. This approach is in contrast with most prior studies, which employ a dyadic view of human-AI interaction and consider either the AI agent and the user [33] or an IS artifact and its creator [27,32,74]. Viewing AI alignment through the lens of the agentic IS framework reveals three possible alignments, all of which are potentially relevant to trust transference in the context of AI agents: creator-AI alignment, user-AI alignment, and AI internal-alignment. Next, we discuss how each of these three possible alignments may be achieved.

First, creators are continuously exploring different approaches to aligning AI agents with their goals and values. However, due to the complexity and probabilistic nature of modern AI agents [6], achieving perfect alignment with their goals and values is a difficult task [45]. This challenge has prompted significant investments in the industry. For instance, one of OpenAI's objectives in developing GPT-4 was to create an instruction-tuned large language model (LLM) that could better align the agent's behavior with their values and avoid crossing ethical red lines [70]. It is therefore imperative to examine how users perceive an AI agent's ability to align with the goals and values of its creator. In keeping with academic research and industry terminology

[53,70], we refer to this attribute as *creator-based steerability*, which is the perceived ability of an AI agent to be directed to behave in a manner that aligns with its creator’s goals and values.

Second, new technologies used in modern AI agents allow users to steer these agents toward greater alignment with the user’s own goals and values. For example, GPT-4 utilizes few-shot learning (FSL) to enable users to fine-tune the agent’s responses based on their specific requirements and values [70]. By specifying the factors they want the agent to consider in generating responses, users can steer the agent to generate more customized responses [70]. Therefore, it is important to understand how users perceive the agent’s ability to be steered to become aligned with their goals and values. Consistent with academic research and industry terminology [53,70], we refer to this attribute as *user-based steerability*, defined as the perceived ability of an AI agent to be directed to behave in a manner desired by its user.

Third, machine learning scholars and practitioners have explored various ways to increase the autonomy of AI agents, including the use of agentic design patterns [66] and concepts such as memories, planning, and reflection [72]. For instance, Auto-GPT² and BabyAGI³ employ LLMs within numerous loops of reasoning, planning (generating, criticizing, and adjusting), and execution to create AI agents with high degrees of autonomy. While technological, moral, and regulatory limitations currently constrain the full autonomy of these AI agents [19,41], they are capable of some degree of independent decision-making and operation [6,50]. The level of autonomy reflects an agent’s capacity to act independently, without direct input from its creators or users [6]. Autonomous agents can potentially learn from interactions with their environment and operate in ways that are not explicitly designed or steered by humans [e.g., see 62]. Therefore, it is necessary to understand how users perceive an agent’s ability to engage in autonomous decision-making. In line with prior research [6], we refer to this attribute as *autonomy*, defined as the perceived ability of an AI agent to make autonomous choices based on its self-determined objectives.

² <https://github.com/Significant-Gravitas/Auto-GPT>, accessed April 25, 2023

³ <https://github.com/yoheinakajima/babyagi>, accessed April 25, 2023

Taken together, these three attributes of AI agents potentially create an alignment tension that influences trust in AI agents. Based on our review of the literature, how these attributes shape the relationship between users' trust (or lack thereof) in the AI creator and their trust in AI agents is an open theoretical and empirical question. Failing to address this question limits scholarly understanding of users' trust in AI agents. It also limits the ability of managers to make informed decisions on how to design and market their AI agents. Motivated by the theoretical and practical significance of this phenomenon, we pose the following research questions:

RQ1: *What is the relationship between trust in an AI creator and trust in its AI agent?*

RQ2: *What are the impacts of creator-based steerability, user-based steerability, and autonomy on the relationship between trust in an AI creator and trust in its AI agent?*

To answer our research questions, we used randomized experiments as our identification strategy. This approach allowed us to isolate the moderating effect of each of the three AI attributes without concern for omitted variables that are not part of our problem formulation [84]. We conducted four experiments and recruited a total of 1,140 participants. First, we conducted a series of three incentivized 2×2 between-subject factorial design experiments in which we independently manipulated trust in AI creator and one of the three AI attributes to determine their effects on participants' trust in an AI agent. As a robustness check, we then also conducted a $2 \times 2 \times 2 \times 2$ between-subject full factorial design experiment in which we independently manipulated trust in AI creator and each of the three AI attributes in a single experiment.

Our research contributes to the existing body of knowledge in three ways. First, we contribute to the emerging literature on AI alignment by considering the entity with whom the AI's goals and values should be aligned and identifying three attributes of AI agents that influence the formation of trust in AI agents: creator-based steerability, user-based steerability, and autonomy. Second, we contribute to the literature on human-AI interaction by drawing attention to the need to theorize from a triadic view that includes the user, the AI agent, and the creator of the agent rather than the currently predominant dyadic view of human-AI interaction that focuses only on the user and the AI agent. Finally, we contribute to trust transference theory by proposing a

contextualized theory of trust transference for AI agents [40]. By examining creator-based steerability, user-based steerability, and autonomy, insights can be gained into how trust is formed in the context of AI agents. Such insights are valuable for the development of trustworthy AI agents that can adapt to different contexts while maintaining alignment with human goals and values.

THEORETICAL BACKGROUND AND HYPOTHESES

Trust in AI Agents

In accordance with the previous literature, we define an *AI agent* as a system that can receive data from its surrounding environment (e.g., users, database systems, and physical sensors), process the data, autonomously generate results or actions, and improve its decision making through data and experience [6,33]. There are several different widely used conceptualizations of trust across different disciplines [51,61,78], but when it comes to trust in AI agents, one must decide whether to use a human-based [51,96] or technology-based [60] conceptualization. Research suggests that people tend to use a human-based conceptualization when interacting with AI agents [48].⁴ Hence, in line with most prior studies on trust in AI agents [33,48], we use a human-based conceptualization of trust and define *trust* as confident positive expectations regarding another’s conduct [51].

Trust in AI agents has been investigated in different streams of research [for a thorough review of trust in AI, see 33]. Findings from research on trust in recommendation agents suggest that familiarity and perceived personalization [46], humanlike features [76], explanation and transparency [95,101], the type of recommendation agent, the method used to elicit preferences, response time, and recommendation content and format [97] influence trust in recommendation agents. Findings from the algorithm aversion literature indicate that people are less likely to rely

⁴ McKnight et al. [60] initially proposed *trust in a specific technology* for dealing with technological artifacts. They argued that human-based and technology-based conceptualizations of trust mainly differ in terms of the object of dependence and nature of the trustor’s expectations. Regarding the former, a human-based concept encompasses agency in both volitional and non-volitional factors whereas a technology-based concept encompasses non-volitional factors only. Regarding the latter, a human-based concept considers perceived ability, integrity, and benevolence whereas a technology-based concept considers functionality, reliability, and helpfulness. However, in a subsequent study, Lankton, McKnight, and Tripp [48] empirically assessed the appropriateness of such a conceptualization when dealing with technological artifacts that have different levels of human-likeness. They concluded that a technology-based conceptualization should be used for non-humanlike technologies such as spreadsheet software, but that a human-based conceptualization of trust should be used for humanlike technologies such as recommendation agents.

on algorithms than on humans [16], especially for subjective tasks [17] such as recommending jokes [102]. Even when an algorithm and a human make the same mistake, users are more likely to stop relying on advice from the algorithm than from the human [25]. Similarly, people are averse to AI agents making a range of ethical decisions, however, the aversion decreases by limiting the agents to an advisory role [8]. Further, research finds that in contexts like medicine, people are averse to AI recommendations because they feel that AI is not as capable as a human in understanding the uniqueness of their situation [55]. In contrast, findings from the algorithm appreciation literature show an overall appreciation of algorithms by users [9,54,103]. In the absence of information on human versus algorithm performance, for instance, laypeople adhere more to advice from an algorithm than from a person [54] or a group of people [35].

Recent studies speculate that the seemingly contradictory findings regarding algorithm aversion and appreciation have their roots in the complex nature of human-AI interaction [17,42] and recommend further research to examine the unique aspects of AI agents that may influence the nature of human-AI interaction [42]. One of these unique aspects is the relationship between an AI agent and its creator [2]. Prior research on traditional IS artifacts (e.g., websites and applications) suggests that trust in the creator of the artifact (e.g., companies, online vendors, and developers) is a strong predictor of a range of trust-related outcomes, including the intention to use [32], purchase and repurchase intention [27], and information disclosure [86]. However, given the fundamental differences between traditional IS artifacts and AI agents (e.g., autonomy) [6,23,82], further research is needed to determine how the unique attributes of AI agents (e.g., AI alignment) may affect trust transfer from an AI creator to an AI agent [33].

Trust Transfer from AI Creator to AI Agent

Trust transference is a phenomenon whereby a person's trust in one entity influences their trust in another entity, either in the same or in different contexts [94]. For instance, trust transfers between companies and their salespeople [5]. The extent of trust transference depends on how the trustor perceives the association between the two entities [65,88]. People's perception of the

association between the two entities can range from viewing the two entities as a single cohesive unit to seeing them as completely independent entities [88].

Trust transference can also take place in a triad that includes a trustor (e.g., a user), a trustee (e.g., an AI agent), and a third party who is related to the trustor and trustee (e.g., the AI creator) [94]. When a person interacts with a dyad, their perception of the two entities depends on the relationship between the two entities in the dyad [38]. If the relationship is perceived to be positive, the person should perceive both entities as either positive or negative to maintain a cognitive balance. If the relationship is perceived to be negative, the person's perceptions of the two entities should be in opposite directions to maintain a cognitive balance. For instance, if a person likes Microsoft (i.e., a positive relationship with Microsoft) and perceives Microsoft and Apple to be enemies (i.e., a negative relationship between the two companies), the person is likely to form a negative view of Apple. This is because people have a tendency toward cognitive balance states in their relationships with other entities [38,100]. Individuals attempt to understand their environment in a way that prevents contradictions that would lead to cognitive imbalance. For example, if someone does not trust an AI creator, they would be disinclined to use its AI agent, as this would create a contradiction that would lead to a state of cognitive imbalance.

In the context of AI agents, we argue that the relationship between an agent and its creator is normally perceived to be positive (i.e., they are perceived to be aligned). While the AI alignment problem can and does exist (i.e., an AI creator may not be able to perfectly align the AI to behave based on the AI creator's goals and values) [30,36], it is unlikely that a layperson, on average, sees an AI agent and its creator as completely independent entities. In other words, an AI agent could be perceived as an "agent" of its creator because the AI creator would build the AI agent to advance the AI creator's intentions. In designing AI agents, AI creators are in a position to embed (at least to some extent) computational logic that is in line with their own goals and values. This enables AI agents to serve as an extension of their creator. As such, a user is likely to perceive an AI creator and its AI agent to be aligned. A trustworthy AI creator is

likely to create an AI agent that can be relied upon. Thus, if a user trusts the AI creator, they are more likely to trust the AI agent. If a user does not trust the AI creator, they are more likely to think that the AI creator has built an AI agent that may give biased recommendations that would not be in the best interest of users [98]. Consistent with trust transference [89], we contend that users transfer their trust from an AI creator to its AI agent and state the following hypothesis:

Hypothesis 1: Trust in AI creator increases trust in an AI agent.

AI Alignment

AI alignment entails ensuring that AI is properly aligned with human goals and values [80]. Empirical work has shown that AI alignment is of great concern to the general public [104]. Given the influence of AI agents on people’s livelihood, the United Nations and many AI practitioners and scholars have raised concerns about the limited research on AI alignment [29] and have called for more work on how AI can be aligned with “shared global values” [92:63].

AI alignment is often referred to as value alignment [30]. In the psychology literature, values refer to cognitive representations of desirable, abstract goals and are different than specific goals in that values are transsituational [79]. However, in the context of AI alignment, “the notion of ‘value’ can serve as a placeholder for many things” [30:417]. Previous literature on AI alignment has used the terms values, goals, desires, intentions, requirements, preferences, and interests to delineate the object of alignment efforts [30,37].

One reason why the AI alignment literature uses ‘value’ as a placeholder is that determining and operationalizing values to effectively align an AI agent with a human stakeholder is challenging [30]. First, individuals may hold very different values from one another, update their values over time, have conflicting values within their own value systems, or have different selves (e.g., want-self versus should-self) [3,4]. Second, even when the values are determined, operationalizing the values in a specific context requires ensuring that the AI agent behaves (in terms of both the outcome and how it produces the outcome) in line with those values. Consequently, it is hard to understand *what* values should be encoded and *how* to encode these values in an AI agent such that it is perceived to be aligned with users’ values. In contrast,

it may be more practical to align an AI agent with human goals, as goals are more concrete and context-specific compared to values.⁵

According to Gabriel [30], AI alignment involves two main aspects, both of which are required in any AI alignment effort: technical and normative. The technical aspect focuses on *how* to encode goals and values in AI agents, whereas the normative aspect focuses on *what* goals and values to encode in AI agents [30]. The technical aspect of AI alignment has been at the heart of recent research in many subfields of machine learning including reinforcement learning [45] and large language models [70,71]. The normative aspect of AI alignment has been alluded to in studies on machine morality [8,56], AI ethics [19], interpretability [11], explainability [77], and transparency [101] and recently received direct attention from the philosophical and regulatory points of view [20,29].

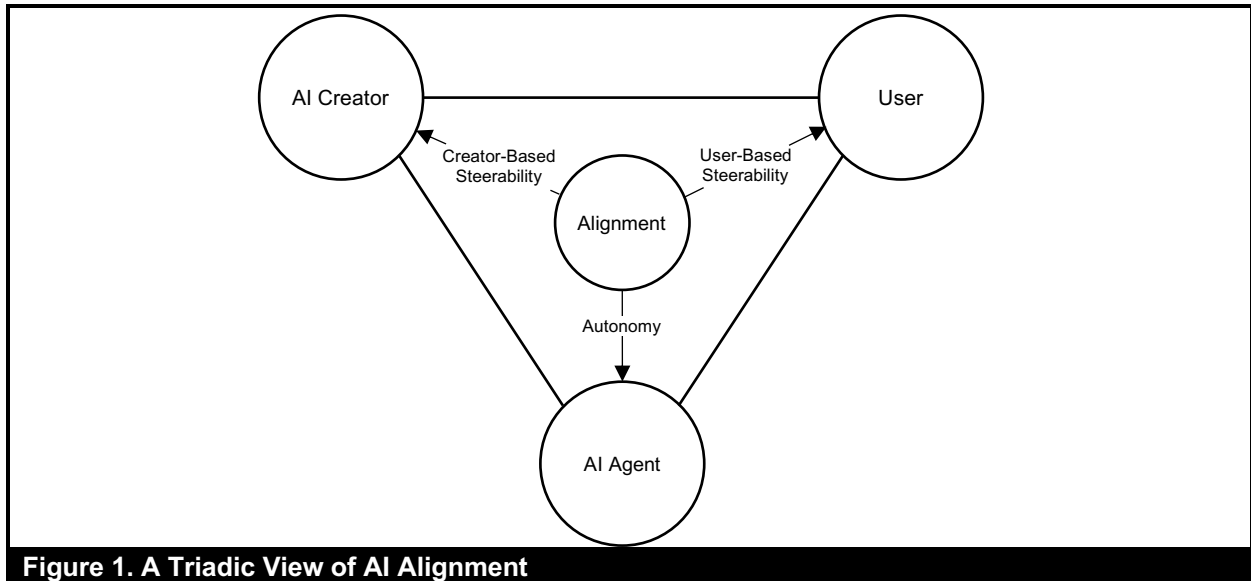
The concept of AI alignment is more meaningful when the AI agent’s fulfilment of a request involves some level of judgment about how it should be fulfilled [30]. For example, a multimodal LLM’s (e.g., Google Gemini) response to the prompt “an image of a successful CEO” could be to generate an image of a white male. This may or may not align with the creator’s or user’s goals and values.

Despite interesting discussions in the literature about normative aspects of AI alignment (e.g., whether the agent should be aligned with expressed intentions versus implied intentions [30]), the target of the alignment has received little attention [31]. Many studies on AI alignment assume that human goals and values are homogeneous across stakeholders [3]. Thus, previous academic work has not considered the need to align an AI agent with multiple stakeholders who have different goals and values.

In the current study, we draw from the agentic IS framework [2] to systematically study the AI alignment problem. The agentic IS framework [2] yields a triadic view of human-AI

⁵ It is important to note that AI alignment does not necessitate making AI agents “follow instructions in an extremely literal way” [30:417]. Alignment instructions may involve revealing a human user’s intentions, preferences, interests, goals, and values instead of providing detailed, mechanical steps to solve the problem (e.g., computer programming code).

interaction, which comprises the AI creator, the user, and the AI agent. This leads to three possible alignments to be considered: creator-AI alignment, user-AI alignment, and internal-alignment (see Figure 1). Internal-alignment does not imply that an AI agent is a stakeholder on its own. Instead, it highlights that AI agents have some level of behavioral independence (based on their autonomy, as discussed later), which is crucial in understanding the AI alignment phenomenon. For instance, an AI agent could “learn internally-represented goals which generalize beyond their training distributions, and pursue those goals” [67:1] or set internal objectives to fulfill a required task and deliver the task through a process that is not deemed appropriate by its external stakeholders [67].



We now elaborate on the three possible alignments to be considered and explain how each influences trust transference.

Creator-AI Alignment: Creator-Based Steerability

In the context of our research, *AI creator* refers to the company associated with creating an AI agent. In practice, an AI creator may be the original developer of an AI agent or the deployer of an AI agent developed by a third party. In both cases, the AI creator is in the position to align the AI agent with its goals and values—albeit not perfectly [45].

In the former case, the AI creator can choose the model type, model architecture, hyperparameters, data preprocessing procedure, training and validation data, and the objective/loss function of the machine learning models that govern the agent’s behavior [34]. An example is OpenAI, which could decide how to design ChatGPT and further align it with their goals and values through reinforcement learning from human feedback (RLHF) [70]. In the latter case, the AI creator can fine-tune the AI agent to be aligned with their goals and values. For instance, a company that chooses to create an AI agent based on GPT-4 as a foundation model can fine-tune the AI agent on their custom data and improve the model’s “truthfulness” to be aligned with their values [57,69]. Therefore, AI creators can attempt to steer their AI agents to reflect their goals and values both pre- and post-deployment.

However, as AI agents become more complicated and take advantage of more complex machine learning algorithms, steering an AI agent to be completely aligned with the AI creator’s goals and values becomes challenging [30,36]. Therefore, creator-AI alignment can theoretically range from completely aligned to completely misaligned [45].

In practice, users may not be able to directly observe how an AI agent is developed, fine-tuned, or maintained by its creator. Thus, their perceptions about creator-based steerability mostly come from impressions based on indirect sources about how the agent works. These sources, which include product descriptions, reviews, advertisements, and news stories, can shape perceptions about the AI agent’s design philosophy, ethical guidelines, and the intentions behind its development, even before any direct interaction. When the AI agent is described as having been created using the company’s own proprietary algorithms, for example, a user might perceive higher creator-based steerability and thus a stronger alignment between an AI agent and its creator (e.g., Google and Imagen [81]). For instance, in 2024, media reports surfaced regarding Google’s efforts to promote diversity and inclusion through their AI agent, Google Gemini [91]. Reportedly, many users were outraged by the very fact that Google attempted to steer the agent toward a specific agenda. In contrast, when the AI agent is described as having been created using algorithms developed by others (e.g., standard, open-source algorithms), a

user might perceive a lower creator-based steerability and thus a weaker alignment between an AI agent and its creator (e.g., thumbsnap.com and its image generator AI based on Stable Diffusion⁶).

We argue that the mere fact that the design of an AI agent allows its creator to embed its own goals and values in the agent decreases trust in the agent (independent of users' trust in the creator). This is predicated on the idea that when an AI agent's design allows its creator to infuse their own goals and values into the agent, it engenders uncertainty and skepticism and thereby undermines trust. First, when interacting with any system, users rely on their ability to anticipate the system's reactions to determine its trustworthiness [96,98]. The notion of creator-based steerability introduces an element of unpredictability and uncertainty. If an AI agent's goals and values can be easily influenced or manipulated by its creator post-deployment, it becomes challenging for users to anticipate its actions. This uncertainty can erode trust as humans are generally averse to uncertainty [33,51].

Second, the complex role of the AI creator, tasked with balancing the needs of numerous stakeholders, presents a potential conflict of interest. If the creator can steer the agent, they may not always prioritize the users' best interest [1]. This concern is amplified by consumer advocates who argue that firms may maintain intentional corporate secrecy regarding their algorithms' inner workings to conceal violation of regulations, discriminatory practices, and consumer manipulation [73]. In other words, opacity regarding the inner workings of AI agents allows firms to covertly enact their goals and values [15]. Thus, the mere presence of a mechanism through which the AI can directly inherit its goals and values from an entity other than its end user raises legitimate concerns about possible future malicious behavior.

In summary, while creator-based steerability offers AI creators a powerful tool to ensure alignment with their goals and values, it also allows for external influence that might not always be aligned with the user's best interests. The mere knowledge that an AI agent can be steered

⁶ <https://github.com/CompVis/stable-diffusion>, accessed October 9, 2022

pre- and post-deployment to reflect the changing goals and values of its creator can be a source of uncertainty and decreased trust in the agent among users. Therefore, we hypothesize that:

Hypothesis 2: Creator-based steerability decreases trust in an AI agent.

As previous research has suggested [89] and as we argued in Hypothesis 1, the extent to which users trust the creator of an AI agent may transfer to their trust in the agent itself. However, we propose that this effect is contingent upon the degree of creator-based steerability exhibited by the AI agent. The rationale is that creator-based steerability affects the salience of the positive relationship between trust in AI creator and trust in the AI agent.

In essence, creator-based steerability serves as a factor that illuminates the nature of the internal relationship between the AI agent and its creator. Particularly, when a user perceives an AI agent to have a high degree of creator-based steerability, it enables them to infer that the creator and the agent are internally related. This, in turn, leads to the perception of AI creator and AI agent as a cohesive unit rather than completely independent entities [88]. This notion is supported by previous research on perceived cohesion and entitativity, where individuals conflate entities that seem deeply intertwined [22]. Consequently, increased creator-based steerability solidifies the perception of an AI agent as an extension of the AI creator.

In contrast, indications that suggest limited creator-based steerability signal some independence of the AI agent from the creator. This perception of separation can attenuate the cognitive association between the two entities, diluting the strength of trust transference. In other words, for a user of a creator-steerable AI agent, their trust in the AI creator is crucial to their trust in the AI agent, as the AI creator can steer the agent to be more congruent with the creator's goals and values. Thus, we state the following hypothesis:

Hypothesis 3: The positive effect of trust in AI creator on trust in an AI agent is stronger for an AI agent with high creator-based steerability than an AI agent with low creator-based steerability.

User-AI Alignment: User-Based Steerability

AI agents may also have the capability to be steered by their users. While not all AI agents are designed to be steered by users, the number of user-steerable AI agents is growing.

Traditionally, most machine learning models are trained before implementation and retrained when enough new data become available [75]. In this approach, the model would be trained for all users at once, and each user's input would not necessarily change the way the agent works for that specific user. However, recent developments in machine learning are changing the user's level of control over the AI agents' behaviors. Particularly, some transfer learning methods⁷, such as one-shot and few-shot learning [99], allow AI agents to be fine-tuned for specific users and problem domains using very few inputs from the user. For instance, GPT-4 allows users to steer the agent to behave according to their requirements and values through system prompts and few-shot learning [for more technical details, see 63]. For example, a user who needs ChatGPT to improve the coherence of a paragraph can steer the agent by adding the following sentence to their prompt: *“Act as if you are an Information Systems researcher who works on trust in AI. When editing a paragraph, do not use overly formal language and do not change the meaning of the text.”* Thus, user-based steerability is possible from a technical perspective and is present in many of today's AI agents. However, it is also important to understand how users *perceive* the user-based steerability of AI agents.

The perception of user-based steerability can be formed based on a variety of interactive means. These can include, but are not limited to, adjusting settings or preferences, using specialized commands or syntax to guide the AI agent's outputs, and engaging in meta-dialogue with the AI to refine its understanding of tasks [1]. Additionally, user engagement with community forums, user guides, fact sheets, and educational materials about the AI agent (e.g., on prompt engineering) can also enhance the perception of user-based steerability by equipping users with the knowledge to guide the AI agent's behavior more effectively [1].

⁷ Transfer learning refers to the retraining of a portion of an already trained model to more appropriately transfer the already learned patterns from one domain to another related domain.

We argue that the user-based steerability of an AI agent empowers users to assert soft controls on the behavior of the agent. Soft control mechanisms influence the agent's behavior by creating shared goals and values [21]. For example, users can steer a content provider AI agent by specifying that they do not want to see specific types of content that might go against their personal values. Therefore, the ability to steer an AI agent allows users to embed their own goals and values into the agent. We posit that user-based steerability can lead to a perceived potential for value congruency between the user and AI agent. Such value congruence enhances users' perception of the AI agent's adherence to principles they deem acceptable. In other words, user-based steerability can enhance the AI agent's perceived integrity. Furthermore, prior literature suggests that value congruence can increase the likelihood of users perceiving the agent as benevolent—acting in the best interest of its user [1,85]. In conclusion, user-based steerability can influence trust in an AI agent by bolstering users' perception of its integrity and benevolence. Thus, we state the following hypothesis:

Hypothesis 4: User-based steerability increases trust in an AI agent.

When a user can steer the AI agent, the AI agent's behavior is more likely to be in line with the user's interests because the user has some level of control over the AI agent's behavior. Therefore, user-based steerability strengthens the user-AI agent relationship leading the user to perceive a positive association with the AI agent. In other words, the user might view the AI agent as an extension of themselves (i.e., as an agent that does things on behalf of its user).

Assuming that the user has a positive relationship with the user-steerable AI agent (as discussed above), trust transference theory suggests that if the user does not trust the AI creator (i.e., if the user has a negative relationship with the AI creator), then they must believe that the relationship between the user-steerable AI agent and its creator is weak. Otherwise, there will be a cognitive imbalance in the user's mind [88,100]. Accordingly, user-based steerability can shift a user's perceived relationship of the AI creator-AI agent from a cohesive dyad toward two independent entities. This shift decouples the perceived link between the AI agent and its creator such that trust transference is weaker. Thus, the user is less likely to perceive that the behavior of

the AI agent is driven by its creator. In other words, for a user of a user-steerable AI agent, their trust in the AI creator becomes less relevant to their trust in the AI agent, as they can steer the agent to be more congruent with their own goals and values. However, for an AI agent with low user-based steerability, users’ trust in the AI creator drives their trust in the AI agent. We therefore pose the following hypothesis:

Hypothesis 5: The positive effect of trust in AI creator on trust in an AI agent is weaker for an AI agent with high user-based steerability than an AI agent with low user-based steerability.

Internal-Alignment: Autonomy

Autonomy is considered to be one of the defining attributes of AI agents [6]. The degree of autonomy denotes an agent’s independence in decision-making and operation [6]. Due to technological, moral, and regulatory limitations, current AI agents vary in their degree of autonomy (fully autonomous AI agents have yet to be developed) [19,41].

An AI agent’s autonomy reflects the agent’s alignment with objectives that are inferred or constructed by the AI agent based on contextual requirements [6]. Autonomous AI agents can sense the environment and set appropriate objectives to be achieved based on what they have learned from previous data.⁸ They can often provide reasoning for their choices (e.g., in the case of explainable AI agents) and continuously monitor their environment to determine progress toward their objectives [47].

An autonomous AI agent can behave based on a set of values or rules that might be unfamiliar or unknown to the users. This unfamiliarity can lead to a sense of anxiety and uncertainty [68], which could be detrimental to people’s trust in autonomous AI agents [24]. This may be exacerbated by negative views of autonomous AI agents propagated in popular culture artifacts. According to Broadbent et al., [12] popular culture artifacts (e.g., podcasts and movies) serve as important contributors to people’s perception of AI agents, and likely decrease trust in

⁸ For a detailed technical discussion on autonomous AI agents, see <https://youtu.be/VRzvpV9DZ8Y?t=808> (streamed September 27, 2022) or <https://openreview.net/pdf?id=BZ5a1r-kVsf> (Version 0.9.2, June 27, 2022) by Yann LeCun.

AI agents by illustrating that such agents could harm humans. While such perceptions might not necessarily be based on concrete facts, scholars have claimed that a negative view of autonomous AI agents exists among the public [90]. Therefore, we hypothesize that:

Hypothesis 6: Autonomy decreases trust in an AI agent.

Further, people typically believe that an autonomous agent is responsible for its own actions [28]. Thus, we postulate that when users perceive a high degree of autonomy in an AI agent, they see the agent as an independent entity—i.e., not part of a cohesive AI creator-AI agent dyad. In this case, the AI agent is perceived to act independently of its creator. Therefore, perceptions about the intentions of the AI creator are less likely to be transferred to the agent. Thus, we suggest the following hypothesis:

Hypothesis 7: The positive effect of trust in AI creator on trust in an AI agent is weaker for an AI agent with a high degree of autonomy than an AI agent with a low degree of autonomy.

Figure 2 shows our research model, which explains trust transference in the context of AI agents based on trust in AI creator, creator-based steerability, user-based steerability, and autonomy and the two-way interactions of trust in AI creator with each of the three moderators.

Table 1 provides a summary of the constructs used in our research model.

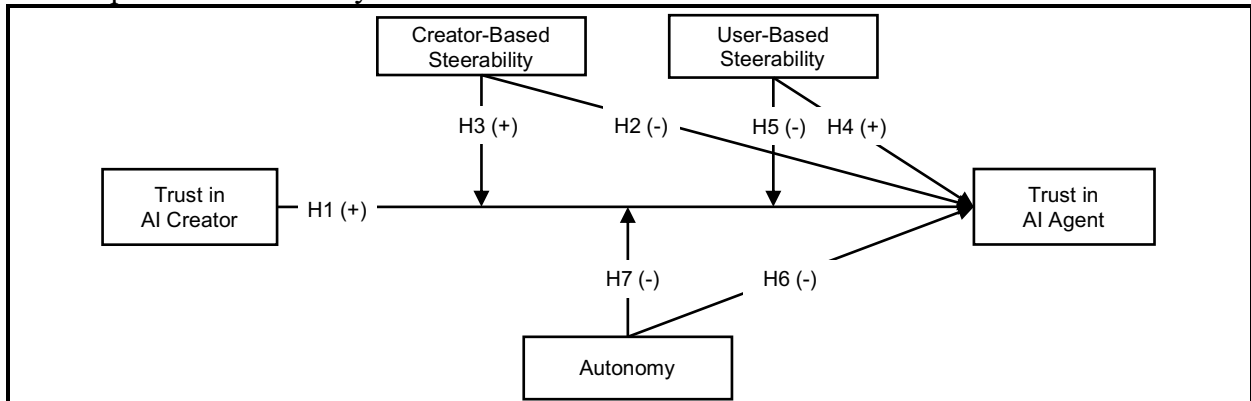


Figure 2. Research Model

In three experiments described in the next section, we independently test key portions of our research model, controlling for age, sex, ethnicity, education, task familiarity, context familiarity, and experience using AI agents, which can potentially influence trust in AI agents [17].

Table 1. Constructs, Definitions, and Examples		
Construct	Definition	Example
Trust	Confident positive expectations regarding another's conduct [52].	An HR manager consistently relies on an AI-based recruitment agent, believing it will fairly and effectively screen candidates despite the inherent risks in HR decision-making tasks.
Creator-Based Steerability	The perceived ability of an AI agent to be directed to behave in a manner that aligns with its creator's goals and values.	A graphic designer reads a news article detailing how the creators of a text-to-image AI agent have designed the agent to promote the creator's values, such as diversity and equity.
User-Based Steerability	The perceived ability of an AI agent to be directed to behave in a manner desired by its user.	A political analyst learns that they can leverage prompting to make an LLM-based AI agent adopt a liberal perspective when summarizing a report.
Autonomy	The perceived ability of an AI agent to make autonomous choices based on its self-determined objectives.	An investor learns that a financial advisor AI agent has the capability to autonomously adjust investment portfolios based on real-time market data and trends without needing explicit approval for each transaction.

METHODOLOGY⁹

Experiments 1-3

As each of the hypothesized interactions in our research model only involves trust in AI creator and one of the AI attributes, we can accurately identify the causal relationships by conducting three separate randomized experiments. Thus, we conducted three incentivized 2×2 between-subject factorial design experiments to manipulate trust in creator and one of the AI attributes independently (experiment 1: trust in creator × creator-based steerability; experiment 2: trust in creator × user-based steerability; experiment 3: trust in creator × autonomy). The Online Appendix provides an overview of the three experiments. To increase participants' engagement, psychological realism, and ecological validity [7], we developed an interactive AI agent through which we delivered our manipulations of AI attributes.

We estimated the number of participants needed for the three experiments using G*Power 3.1.9.6. Based on pilot studies¹⁰, we expected Cohen's $f \approx 0.295$, $f \approx 0.274$, and $f \approx 0.253$ in experiments 1, 2, and 3, respectively. For these effect sizes, $\alpha = 0.05$, power = 0.80, numerator df = 3 (two main effects and one interaction), and the number of groups = 4, we needed 130, 149, and 175 participants in experiments 1, 2, and 3, respectively. Since the effect

⁹ All main experiments (Experiment 1 to 4) were conducted in 2020 and 2021. The additional experiment presented in Online Appendix D was conducted in 2023.

¹⁰ To develop the experimental instruments, we conducted four pilot studies for experiment 1 ($N_1=60$, $N_2=60$, $N_3=80$, $N_4=80$), two pilot studies for experiment 2 ($N_5=80$, $N_6=80$), and five pilot studies for experiment 3 ($N_7=240$, $N_8=80$, $N_9=200$, $N_{10}=80$, $N_{11}=80$). We followed different goals in each of these pilot studies including testing manipulations, adjusting the technical design of the AI agent, and running a small version of the final experiments to estimate the required sample sizes. None of the 1,120 observations collected in the pilot studies were used in our final results.

sizes were not guaranteed, and some participants might fail the attention check question, we chose to recruit 160, 180, and 200 participants in experiments 1, 2, and 3, respectively.

Participants

We recruited all of our participants from Cloud Research, an online participant recruitment platform. Table 2 shows the number of participants in each experiment, the number of participants who passed our attention check question¹¹, their demographics, and the compensation they received.

Table 2. Participants in Experiments 1, 2, and 3												
Exp#	N	Retained N ^(a)	Sex ^(b)			Age			Education	Experience ^(c)	Time Spent ^(d)	Compensation ^(e)
			F	M	O	Min	Mean	Max	Median	Median	Median	Mean
1	160	151	72	79	0	18	38.5	69	4-year college	once-a-week	13.9 minutes	\$1 + \$1
2	180	155	67	86	2	19	38.5	77	4-year college	once-a-week	12.7 minutes	\$1 + \$1
3	200	174	93	81	0	21	42.3	74	4-year college	once-a-week	14.9 minutes	\$1 + \$1
a. Retained N indicates the number of participants who passed the attention check question and were retained for subsequent analyses. b. F: Female, M: Male, O: Other c. Experience: Experience using digital assistants such as Amazon Alexa, Apple Siri, etc. d. Time Spent: The amount of time participants took to complete the experiment. e. \$1 + \$1: \$1.00 base plus a \$1.00 performance-based bonus, which was ultimately paid to all participants in accordance with the IRB.												

Experimental Context and AI Agent

We chose human resource hiring as our experimental context as AI agents (often steerable and autonomous) have been widely used in this context [13]. In each experiment, participants were given the task of hiring a programmer¹² from a pool of ten candidates. The programmer was being hired to develop an app for a fitness company. Since the majority of the available candidates met the job’s minimum requirements, the task did not have an objectively optimal solution (see Figure 3). We conducted a separate study involving 300 participants to ensure that the majority of the candidate resumes would be perceived as a viable fit for the job posting. See the Online Appendix for more details.

¹¹ Following Dietvorst and Bharti [24], we asked the following question to assess participants’ attention: “Your experience with AI agents is important for this survey. In order to demonstrate that you have read the questions carefully, please select other and type the word shoe as your answer to the question below. How often do you use AI agents?” However, unlike Dietvorst and Bharti [24], we asked this question closer to the end of the experiment and still fully compensated the participants who failed to answer it properly. We also recorded a log of each participant’s clicks on the screen in the hiring task to make sure they paid an acceptable level of attention to the task.

¹² Our data show that 137 out of 151, 142 out of 155, and 153 out of 174 of the participants in experiments 1, 2, and 3 had some familiarity with the hiring task and 115 out of 151, 119 out of 155, and 122 out of 174 in experiments 1, 2, and 3 had some familiarity with programming. Nonetheless, we included familiarity with programming and familiarity with hiring as control variables in our empirical models.

We developed an interactive agent in JavaScript and integrated it with the rest of our study through the Qualtrics XM platform’s APIs. The agent leveraged Google’s BERT model¹³ to calculate the best matches for a given job based on the job description and candidates’ résumés.

Overall Procedure

In each experiment, participants played the role of a company’s employee who was in charge of the hiring decision. As an incentive, participants were told that they would receive a \$1 bonus if their hire turned out to be among the top 3 best performers based on candidates’ real performance data in similar positions. We ensured that participants understood their roles by asking them to type in their roles and the bonus structure in a text box. Next, participants were told that they could use an AI agent developed by a company named NextGen to help them choose a candidate. Then, they were shown a news article about NextGen that was read to them by a newscaster’s voice. Afterward, a single AI attribute was manipulated (as discussed in more detail later). Participants were then presented with a job description, candidates’ résumés, and an interactive AI agent (see [Video 1](#)) (see Figure 3).

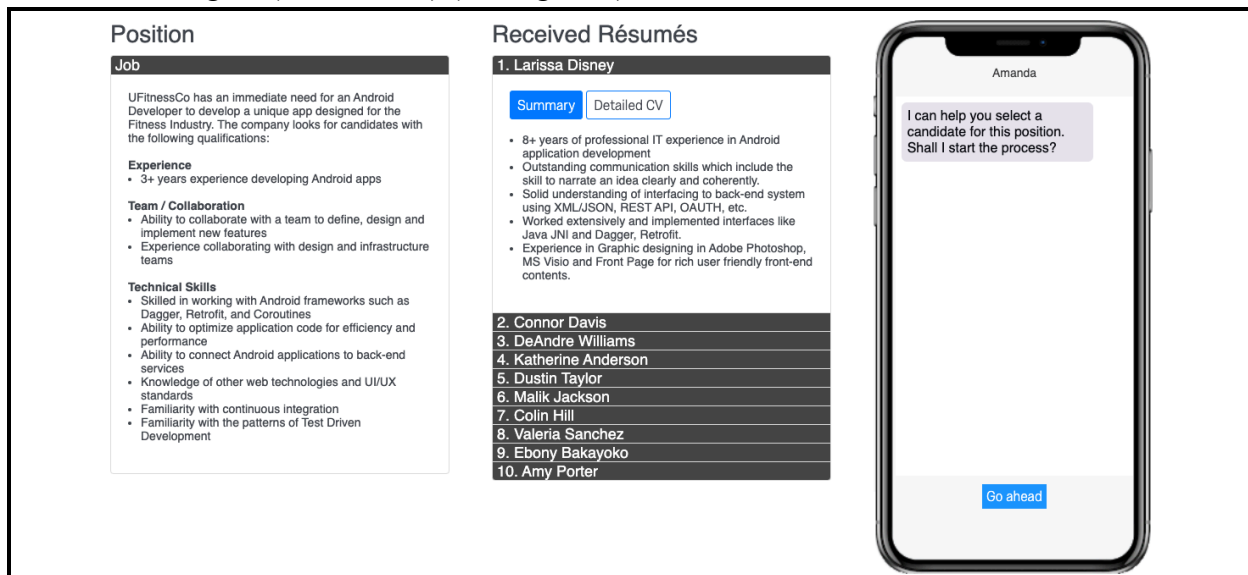


Figure 3. Hiring Task

Participants could see a summary of each candidate’s résumé by expanding the tab for that candidate. The summaries included the most relevant skills related to the job description to

¹³ <https://tfhub.dev/google/collections/bert/1>, accessed October 10, 2021

simplify the hiring task for the participants with less technical knowledge. Participants could also view the complete version of the résumés without leaving the experiment page. The ten candidates and their résumés were based upon a list of actual candidates found on a reputable IT personnel recruitment website (hireITpeople.com). Next to the job description and résumés was an interactive AI agent that participants could use to help them choose a candidate. The agent would first shortlist three candidates and then furnish one candidate as the final recommendation (see Figure 4). The agent was set to recommend the same candidate (candidate number five) across Experiments 1 to 3 and all conditions to avoid confounding the results with other factors that could have influenced our results such as the effect of stereotypically racial names that may influence participants' trust. By keeping the recommendation constant across all conditions, we ensured that the observed differences between conditions would not be due to the specific résumé or candidate's name.



Figure 4. User-AI Agent Interaction During the Hiring Task

After the participant used the AI agent, a question automatically appeared at the bottom of the screen, asking the participant to specify their choice (i.e., the candidate to be hired). To ensure that the participant understood the process, we created an on-screen interactive tutorial that highlighted and explained each section of the screen before the participant started the hiring task. After the participants made their choices, they were asked to answer a series of questions. In the end, in line with our IRB approved protocol, we told all participants that they would

receive the complete bonus regardless of their choices. To see **larger screenshots**, see Online Appendix. To see **video demonstrations** of the experiments, please visit [Video Playlist](#).

Trust in AI Creator Manipulation. In each experiment, we manipulated trust in AI creator by changing the news article about the AI creator. In each condition, we included sentences that could increase or decrease trust in the AI creator [98]. Figure 5 shows the two vignettes used to manipulate trust in AI creator.

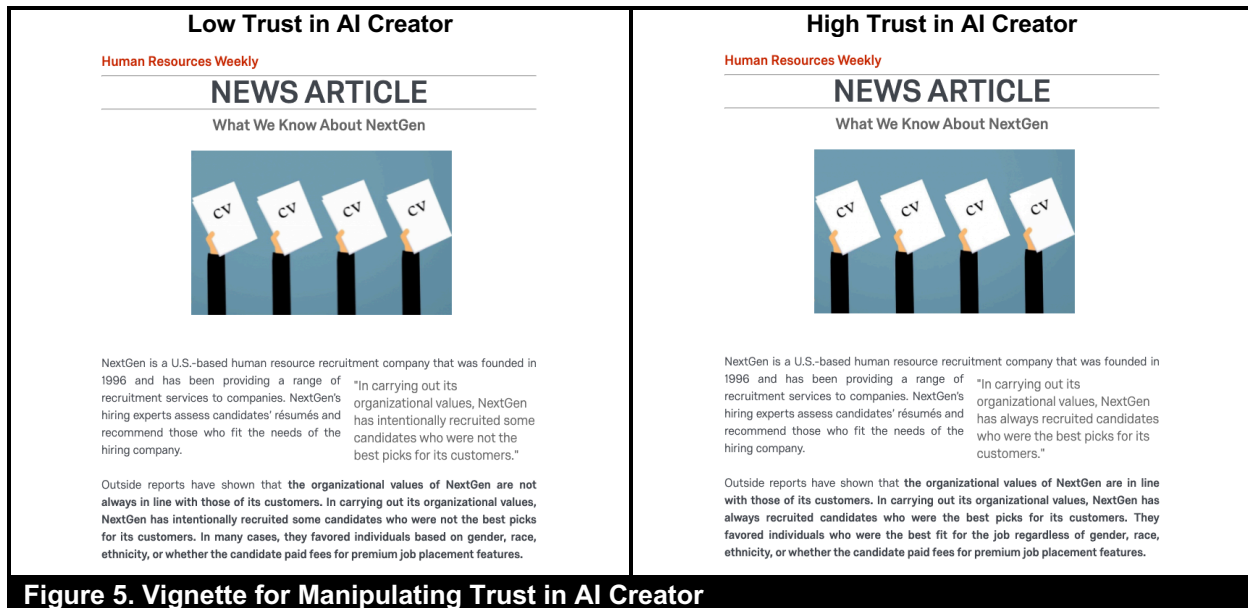


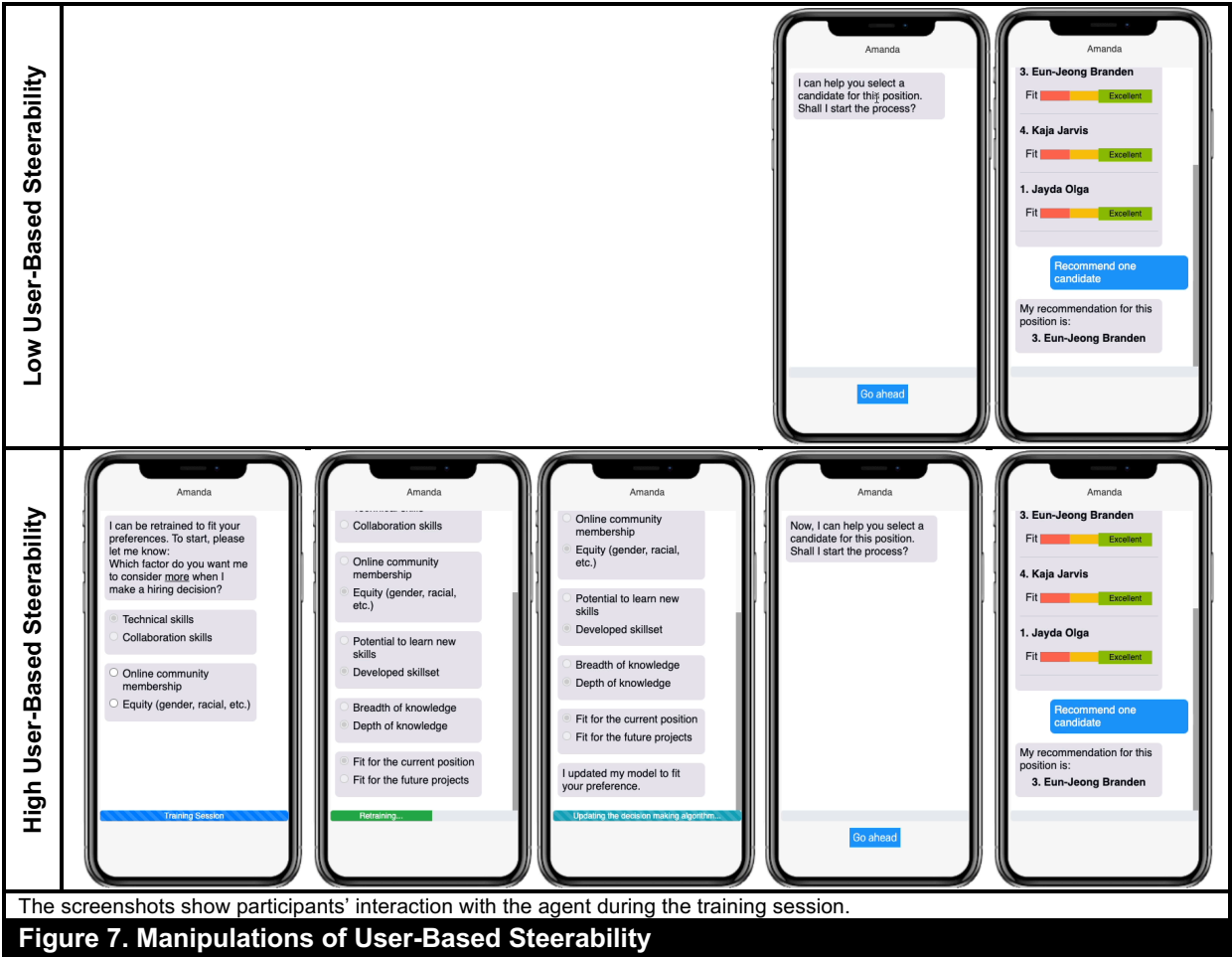
Figure 5. Vignette for Manipulating Trust in AI Creator

Creator-Based Steerability Manipulation. After seeing the news article about the AI creator and before beginning the hiring task, participants were shown a news article about Amanda, a recruitment AI agent recently added to NextGen’s hiring services. The news article about the AI agent, which was read to participants with a newscaster’s voice, described the agent as using either an industry-wide standard algorithm, development over which the AI creator had little direct control (low creator-based steerability), or a proprietary algorithm, development over which the AI creator had complete control (high creator-based steerability). We further reinforced these manipulations in the hiring task by adding a label that specified whether the agent was using an industry-standard algorithm (see [Video 2](#)) or NextGen’s proprietary algorithm (see [Video 3](#)) (see Figure 6).



User-Based Steerability Manipulation. After seeing the news article about the AI creator and before beginning the hiring task, we asked participants to try the AI agent. Specifically, we gave participants a very simple job description alongside five simplified résumés. We asked them to try the AI agent to see how it could help them choose among the candidates. As shown in Figure 7, participants in the low user-based steerability condition directly used the agent to receive a recommendation (see [Video 4](#)). However, participants in the high user-based steerability condition first trained the agent by specifying five factors that they wanted the agent to consider in choosing the candidate and then used the agent to receive a recommendation (see [Video 5](#)). For instance, a participant can steer the AI agent to be aligned with their goals and values by specifying that the agent should consider equity in its hiring decisions. The process of training resembled few-shot learning (FSL), which is a common

approach used to increase user-based steerability in AI agents [70]. However, to keep the conditions comparable, in this trial task and the actual hiring task, the agent gave all participants the same recommendation regardless of the user-based steerability condition.



Autonomy Manipulation. After seeing the news article about the AI creator and before the hiring task, participants were shown a news article about Amanda that was read to them with a newscaster’s voice. The news article mentioned that due to its algorithms, Amanda was either not capable of (low autonomy condition) or capable of (high autonomy condition) making autonomous choices based on its self-determined objectives. We further reinforced the manipulation of high autonomy in the hiring task by adding a label that specified the agent was using the “Autonomous Choice-Making Engine” (see [Video 6](#)) (see Figure 8).



Figure 8. Manipulations of Autonomy

Measures

In line with previous research on AI agents, we measured trust in AI agent as a binary variable indicating whether or not the participant chose the same candidate that the AI agent recommended—i.e., a behavioral proxy for trust in AI agent [16].¹⁴ If the participant chose that candidate, the dependent variable is 1, if they chose any of the other nine candidates, the dependent variable is 0. We used a post-test questionnaire to measure the effectiveness of the manipulations (see Online Appendix) and several control variables.

¹⁴ While it is possible that sometimes a participant's choice happens to be the same as the AI's choice, this possibility is statistically the same in all experimental conditions. Therefore, the random assignment of participants to experimental conditions ensures that the observed difference between conditions is due to the treatment.

Results

Manipulation checks indicate that our manipulations were successful. Table 3 shows the details of Cronbach's α of the measurements, the variables' values under the low and high experimental conditions, the difference between low and high conditions, and whether the differences were significant.

We conducted three sets of hierarchical logistic regressions¹⁵, one set per experiment, to estimate the effects of trust in AI creator (low = 0, high = 1), an AI attribute (low = 0, high = 1) (creator-based steerability in Experiment 1, user-based steerability in Experiment 2, and autonomy in Experiment 3), and their interaction on trust in the AI agent, while including our control variables.¹⁶ Table 4 shows the results.

Table 3. Manipulation Checks												
	Experiment 1: AI Attribute: Creator-Based Steerability				Experiment 2: AI Attribute: User-Based Steerability				Experiment 3: AI Attribute: Autonomy			
Variable	α	Low	High	Δ	α	Low	High	Δ	α	Low	High	Δ
Trust in AI Creator	0.989 (3 items)	2.360 (1.239)	5.458 (1.048)	$t(149)$ = 16.582 $p < 0.001$ $d = 2.699$ success: ✓	0.990 (3 items)	2.439 (1.334)	5.489 (1.073)	$t(153)$ = 15.555 $p < 0.001$ $d = 2.503$ success: ✓	0.990 (3 items)	2.434 (1.249)	5.366 (1.058)	$t(172)$ = 16.579 $p < 0.001$ $d = 2.520$ success: ✓
AI Attribute	0.977 (3 items)	2.953 (1.471)	6.169 (0.835)	$t(149)$ = 16.370 $p < 0.001$ $d = 2.666$ success: ✓	0.982 (3 items)	3.346 (1.653)	5.719 (0.952)	$t(153)$ = 10.932 $p < 0.001$ $d = 1.756$ success: ✓	0.949 (5 items)	2.400 (1.420)	4.622 (1.613)	$t(172) = 9.617$ $p < 0.001$ $d = 1.462$ success: ✓
α : Cronbach's α Δ : difference between the "low" and "high" experimental conditions (manipulation check was calculated as the average of manipulation questions) d : Cohen's d Low and High: mean values under low and high experimental conditions (standard deviations in parentheses)												

Main Effects. We found that trust in AI creator increases trust in the AI agent ($\beta_{exp1} = 0.700, p = 0.034; \beta_{exp2} = 1.079, p = 0.005; \beta_{exp3} = 0.943, p = 0.008$), providing support for Hypothesis 1. Specifically, the odds of trusting the AI agent for people in the high trust in AI creator condition is more than twice the odds of trusting the AI agent for those in the low trust in AI creator condition ($OR_{exp1} = 2.014, p = 0.034; OR_{exp2} = 2.943, p = 0.005; OR_{exp3} = 2.568, p = 0.008$).¹⁷

¹⁵ All analyses were run using Stata version 16.1.

¹⁶ All results in Experiments 1, 2, and 3 remain the same in terms of direction and significance with or without the inclusion of the control variables.

¹⁷ OR = odds ratio

Table 4. Logistic Regression Results for Experiments 1, 2, and 3

	Experiment 1 AI Attribute: Creator-Based Steerability			Experiment 2 AI Attribute: User-Based Steerability			Experiment 3 AI Attribute: Autonomy		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Observations	151	151	151	155	155	155	174	174	174
Pseudo R ²	8.19%	10.21%	13.41%	15.09%	18.94%	22.47%	14.47%	18.12%	20.56%
Main Effects & Interactions									
Trust in AI Creator		0.700** (0.385)	-0.213 (0.527)		1.079*** (0.418)	2.341**** (0.665)		0.943*** (0.394)	2.039*** (0.658)
AI Attribute		0.356 (0.395)	-0.631 (0.564)		0.634* (0.412)	1.747*** (0.613)		0.555 (0.420)	1.568** (0.646)
Interaction			1.979*** (0.792)			-2.303*** (0.866)			-1.905** (0.861)
Control Variables									
Constant	-3.006*** (1.120)	-3.471*** (1.171)	-3.126*** (1.196)	-1.759 (1.367)	-1.997 (1.426)	-3.320** (1.558)	-0.180 (1.237)	-0.622 (1.289)	-1.054 (1.342)
Familiarity with Programming	0.019 (0.210)	0.049 (0.212)	0.079 (0.219)	0.331* (0.180)	0.308 (0.191)	0.392* (0.202)	-0.300 (0.192)	-0.264 (0.200)	-0.313 (0.209)
Familiarity with Hiring	-0.313 (0.212)	-0.283 (0.219)	-0.356 (0.226)	-0.189 (0.187)	-0.306 (0.203)	-0.306 (0.213)	-0.138 (0.199)	-0.064 (0.210)	-0.066 (0.219)
Age	0.046** (0.019)	0.044** (0.019)	0.048** (0.020)	-0.006 (0.018)	0.001 (0.019)	0.004 (0.019)	-0.020 (0.018)	-0.024 (0.019)	-0.024 (0.019)
Sex (Female=0)									
Male	0.710* (0.399)	0.673* (0.404)	0.839** (0.423)	1.074** (0.431)	1.153** (0.447)	1.196** (0.467)	0.705* (0.401)	0.617 (0.417)	0.699* (0.425)
Ethnicity (White=0)									
Black or African American	0.378 (0.868)	0.621 (0.877)	0.562 (0.921)	-0.580 (0.797)	-0.650 (0.821)	-0.985 (0.865)	-0.635 (0.829)	-0.724 (0.823)	-0.828 (0.852)
Asian	-0.232 (0.603)	-0.199 (0.626)	-0.170 (0.619)	-0.411 (0.675)	-0.559 (0.705)	-0.753 (0.740)	-0.774 (0.687)	-0.705 (0.690)	-0.675 (0.696)
Latino or Hispanic	-0.759 (0.899)	-0.772 (0.898)	-0.928 (0.904)	2.975** (1.356)	2.262* (1.351)	2.355* (1.396)	-1.027 (1.186)	-0.987 (1.200)	-0.884 (1.234)
Other	-	-	-	-0.923 (1.297)	-1.311 (1.324)	-1.391 (1.347)	-	-	-
Education (High School or Less = 0)									
Some College	0.841 (0.796)	0.797 (0.808)	0.772 (0.822)	-0.506 (0.779)	-1.052 (0.834)	-0.989 (0.880)	-0.126 (0.800)	-0.504 (0.839)	-0.535 (0.853)
2-year College Degree	0.548 (0.881)	0.382 (0.899)	0.171 (0.936)	0.450 (0.856)	-0.096 (0.904)	-0.226 (0.958)	-0.472 (0.861)	-0.857 (0.894)	-0.874 (0.911)
4-year College Degree	0.827 (0.700)	0.796 (0.709)	0.792 (0.727)	0.011 (0.654)	-0.350 (0.684)	-0.452 (0.725)	-0.494 (0.713)	-0.766 (0.744)	-0.771 (0.762)
Master's Degree	1.167 (0.816)	1.127 (0.826)	1.146 (0.851)	-0.917 (0.808)	-1.294 (0.841)	-1.412 (0.882)	-0.243 (0.784)	-0.296 (0.809)	-0.540 (0.849)
Doctorate Degree	-0.025 (1.178)	0.167 (1.227)	-0.097 (1.264)	-0.196 (1.457)	-1.348 (1.478)	-1.133 (1.548)	-1.543 (1.021)	-1.550 (1.033)	-2.014* (1.080)
Past Experience Frequency (At least once a day=0)									
At least once a week	0.353 (0.469)	0.249 (0.479)	0.210 (0.486)	1.088* (0.614)	1.087* (0.630)	1.170* (0.656)	1.451*** (0.559)	1.440** (0.575)	1.520** (0.601)
At least once a month	0.491 (0.705)	0.521 (0.721)	0.822 (0.751)	0.961 (0.793)	0.714 (0.826)	0.780 (0.855)	0.690 (0.694)	0.582 (0.706)	0.480 (0.728)
Never	0.597 (0.749)	0.313 (0.773)	0.207 (0.788)	0.906 (0.846)	1.010 (0.873)	1.432 (0.921)	0.744 (0.854)	0.548 (0.882)	0.493 (0.907)
Past Experience Agent (Not used=0)									
IoT-Alexa	0.383 (0.466)	0.490 (0.475)	0.649 (0.495)	1.254* (0.669)	1.025 (0.702)	1.200 (0.731)	0.881 (0.549)	0.793 (0.565)	0.843 (0.583)
IoT-Google	0.388 (0.550)	0.170 (0.572)	0.066 (0.592)	0.267 (0.690)	0.126 (0.719)	0.326 (0.733)	-0.627 (0.666)	-0.694 (0.682)	-0.629 (0.707)
Phone-Alexa	0.614 (0.609)	0.645 (0.620)	0.654 (0.630)	1.107 (0.865)	1.279 (0.907)	1.463 (0.964)	0.230 (0.656)	0.119 (0.670)	-0.114 (0.694)
Phone-Google	-0.184 (0.508)	-0.277 (0.518)	-0.332 (0.529)	0.592 (0.565)	0.500 (0.574)	0.796 (0.609)	0.800 (0.532)	0.687 (0.558)	0.668 (0.573)
Phone-Siri	0.432 (0.478)	0.361 (0.493)	0.409 (0.507)	-0.036 (0.565)	0.027 (0.585)	0.412 (0.630)	0.943* (0.515)	0.711 (0.531)	0.673 (0.537)
Other	-	-	-	-0.567 (1.226)	-0.344 (1.409)	0.406 (1.468)	-1.241 (1.215)	-1.050 (1.208)	-0.885 (1.192)

- **** p<0.001, *** p<0.01, ** p<0.05, * p<0.1
- One-tailed tests were used for directional hypotheses
- Standard errors in parentheses
- "-" indicates coefficients that were omitted because of a lack of variance in data.
- Interaction: interaction between trust in AI creator and a given AI attribute (i.e., creator-based steerability, user-based steerability, autonomy)
- Variables under "past experience agent" indicate whether participants used each of the agents at least once a week.
- None of the participants self-identified as "American Indian or Alaska Native" or "Native Hawaiian or Pacific Islander." Therefore, these categories were omitted from the results.
- Only two participants self-identified as "other" sex. The dummy variable for this group could not be statistically identified. Therefore, we estimated the model by setting the sex for these two observations to 0. Setting the variable to 1 or removing these two observations did not significantly change our results.

We did not find supporting evidence for the negative effect of creator-based steerability on trust in the AI agent (Hypothesis 2; $\beta = 0.356, p = 0.817$). However, the model provided some support for the positive effect of user-based steerability on trust in the AI agent ($\beta = 0.634, p = 0.062$), providing support for Hypothesis 4. The odds of trusting the AI agent for participants who used an agent with high user-based steerability is about two times the odds of trusting the AI agent for those who used an agent with low user-based steerability ($OR = 1.885, p = 0.062$). On the other hand, our results did not provide supporting evidence for the negative effect of autonomy on trust in the AI agent (Hypothesis 6; $\beta = 0.555, p = 0.907$).

Moderation Effects. We found evidence that creator-based steerability strengthens the positive effect of trust in AI creator on trust in the AI agent ($\beta = 1.979, p = 0.007$), providing evidence in support of Hypothesis 3. Specifically, for an AI agent with high creator-based steerability, the odds of trusting the AI agent for people in the high trust in AI creator condition are nearly 6 times greater than those in the low trust in AI creator condition ($OR = 5.843, p = 0.002$). However, for an AI agent with low creator-based steerability, the odds of trusting the AI agent are not statistically different for participants in low and high trust in AI creator conditions ($OR=0.808, p=0.685$), indicating that low creator-based steerability dampens the negative effect of low trust in AI creator on trust in the AI agent.

We found evidence that user-based steerability weakens the positive effect of trust in AI creator on trust in the AI agent ($\beta = -2.303, p = 0.004$), lending support for Hypothesis 5. Specifically, for an AI agent with low user-based steerability, the odds of trusting the AI agent for people in the high trust in AI creator condition is more than 10 times what we observed for people in the low trust in AI creator condition ($OR = 10.387, p < 0.001$). However, for an AI agent with high user-based steerability, the odds of trusting the AI agent are not statistically different for participants in low and high trust in AI creator conditions ($OR = 1.039, p = 0.946$), indicating that user-based steerability reduces the negative effect of low trust in AI creator on trust in the AI agent.

Finally, we found evidence that autonomy weakens the positive effect of trust in AI creator on trust in the AI agent ($\beta = -1.905, p = 0.014$), thus providing evidence supporting Hypothesis 7. Specifically, with a low autonomy AI agent, the odds of trusting the AI agent for people in the high trust in AI creator condition is more than 7.5 times greater than what we observed for people in the low trust in AI creator condition ($OR = 7.684, p = 0.001$). However, with a high autonomy AI agent, the odds of trusting the AI agent are not statistically different for participants in the low and high trust in AI creator conditions ($OR = 1.144, p = 0.801$), indicating that high autonomy reduces the adverse effect of low trust in AI creator on trust in the AI agent. Figure 9 shows interaction plots allowing us to visualize these results in terms of probabilities of trust in the AI agent.¹⁸

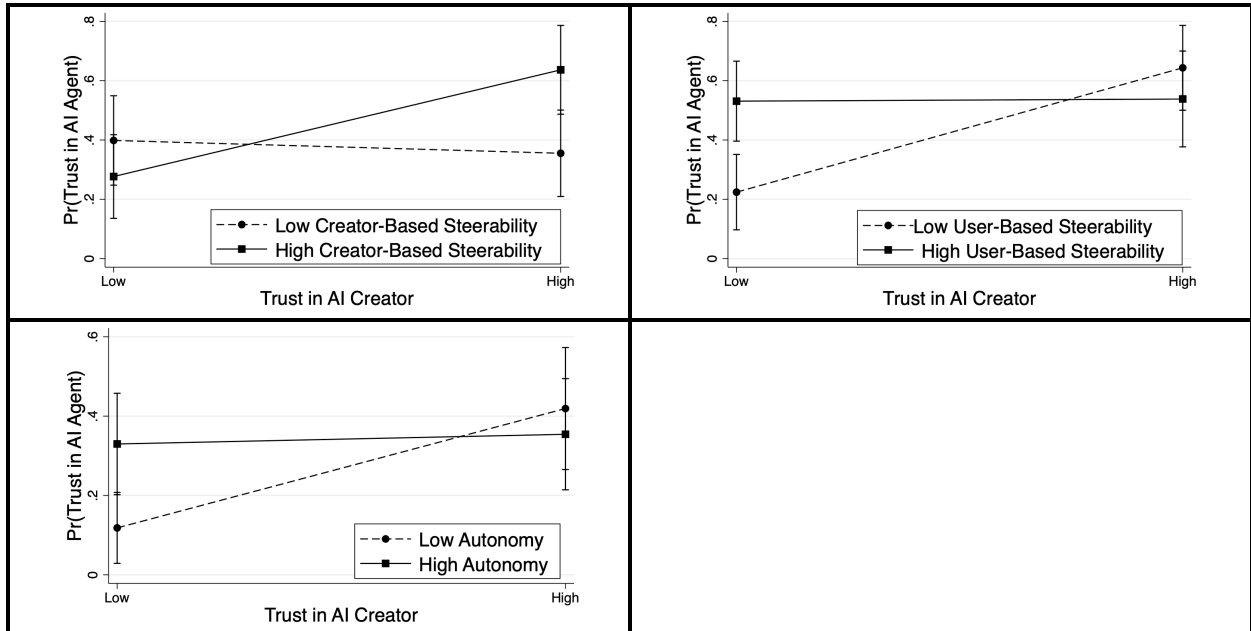


Figure 9. Moderating Role of Creator-Based Steerability, User-Based Steerability, and Autonomy

Experiment 4: Robustness Check

In the first three experiments, we independently examined the theorized relationships among trust in creator, creator-based steerability, user-based steerability, and autonomy. However, our

¹⁸ In our theorization, we argued that the transference of users' trust in creator to their trust in AI agent is moderated by AI attributes. An alternative explanation is that trust in creator may influence the magnitude of users' beliefs about AI attributes. Our analysis, however, reveals that the manipulation of trust in AI creator does not systematically change participants' perception of creator-based steerability (Exp 1: $p=0.888$; Exp 4: $p=0.251$), user-based steerability (Exp 2: $p=0.233$; Exp 4: $p=0.658$), or autonomy (Exp 3: $p=0.473$; Exp 4: $p=0.376$). Thus, it is unlikely that trust in AI creator impacted the magnitude of the beliefs about AI attributes.

approach had a few limitations, which we aimed to address in Experiment 4. First, we were unable to include all three AI attributes in a single experiment due to our experimental designs. Replicating our findings using an experiment in which all factors are manipulated would add robustness to our findings. Second, we measured participants' behavioral trust, ascertaining whether they relied on the AI agent's recommendation, as a behavioral proxy for trust. This is a common approach in fields such as behavioral economics, psychology, and human-AI interaction especially in the context of simulated interactions (e.g., in game theory research [44,52] and AI agents [16,33]). However, the relationship between willingness or intention to rely on an entity and reliance on the entity could be complicated [52,83] and thus replicating our results using a measure of people's trusting intentions (i.e., their willingness to rely on an AI agent) would also add robustness to our findings.

To address these issues, we conducted a $2 \times 2 \times 2 \times 2$ experiment in which we manipulated all independent variables in a single experiment and measured participants' trusting intentions. To do so, we modified our experimental design. Our previous manipulations of AI attributes did not allow us to effectively introduce several AI attributes in a single experiment. For instance, we reinforced our manipulations in the experimental tasks by adding labels associated with a specific AI attribute (see Figure 8). However, the AI agent would appear unrealistic if we were to add multiple labels. Therefore, we simplified the experimental procedure and employed a scenario-based experiment in which the participants were exposed to information about an AI creator and all three attributes of the AI agent created by it.

We estimated the number of participants needed for our study using G*Power 3.1.9.6. Based on pilot studies¹⁹, we expected $f \approx 0.250$. For this effect size, $\alpha = 0.05$, power = 0.80, numerator df = 15 (4 main effects, 6 two-way interactions, 4 three-way interactions, 1 four-way interaction, with all factors having two levels), and the number of groups = 16, we needed at

¹⁹ We conducted four pilot studies with a total of 512 participants to develop the experimental instruments for our study. In the pilot studies, we focused on the content and delivery of the vignettes used for our experimental manipulations, and the required sample size.

least 314 participants in experiment 4. Since the effect size was not guaranteed, and some participants might miss the attention check question, we chose to recruit 600 participants.

Participants

All participants were recruited from Cloud Research and paid \$1. Table 5 indicates the number of participants and their demographics. In this experiment, we used the same attention check question discussed in previous experiments.

Table 5. Participants in Experiment 4										
N	Retained N^(a)	Sex^(b)			Age			Education	Experience^(c)	Time Spent^(d)
		F	M	O	Min	Mean	Max	Median	Median	Median
600	572	244	326	2	20	38.6	78	4-year college	once-a-week	5.7 minutes
a. Retained N indicates the number of participants who passed the attention check question and were retained for subsequent analyses. b. F: Female, M: Male, O: Other c. Experience: Experience using digital assistants such as Amazon Alexa, Apple Siri, etc. d. Time Spent: The amount of time participants took to complete the experiment.										

Procedure

We asked the participants to read a description of NextGen, a fictitious company that creates AI agents. Next, we asked them to fill out a survey about the company (trust in creator). In the next section, we asked them to read a description of Amanda, an AI agent created by NextGen. Then we asked the participants to fill out a survey about the agent. We concluded the experiment by asking demographic questions and debriefing the participants, explaining that the company and the agent were fictitious.

Manipulations. We manipulated trust in creator, creator-based steerability, user-based steerability, and autonomy independently. Trust in creator was manipulated by asking the participants to read a description that induces either low trust or high trust. Figure 10 shows the two vignettes used to manipulate trust in creator.

Creator-based steerability was manipulated by describing the AI agent’s design as allowing (in high creator-based steerability treatment) or not allowing (in low creator-based steerability treatment) its creator to embed its own values in the agent. User-based steerability was manipulated by informing the participants that the agent’s design allows (in high user-based steerability treatment) or does not allow (in low user-based steerability treatment) users to retrain the agent to behave in the way they like. Finally, autonomy was manipulated by stating that the AI agent’s algorithms allow (in high autonomy treatment) or do not allow (in low autonomy

treatment) the agent to make autonomous choices based on its self-determined objectives. Figure 11 provides a summary of these three manipulations. We randomized the order in which the manipulations of AI attributes were presented to participants.



Figure 10. Vignette for Manipulating Trust in Creator

Creator-Based Steerability Manipulation


Autonomy Manipulation

User-Based Steerability Manipulation

Technology Report

NEWS ARTICLE

What We Know About Amanda



Amanda is a digital assistant developed by NextGen corporation. Users can use Amanda to do many tasks such as controlling smart home devices, setting up business meetings, and adding events to their calendars.

A recent customer survey has summarized Amanda's **key features that you should remember:**

- Amanda's design **does not allow NextGen corporation to embed its own values** in the agent.
- Due to its algorithms, Amanda is **capable of making completely autonomous choices** based on its self-determined objectives.
- Amanda's design **does not allow users to retrain** the agent to behave in the way they like.

Construct	Level	Manipulation Content
Creator-Based Steerability	Low	Amanda's design does not allow NextGen to embed its own values in the agent.
	High	Amanda's design allows NextGen to embed its own values in the agent.
User-Based Steerability	Low	Amanda's design does not allow users to retrain the agent to behave in the way they like.
	High	Amanda's design allows users to completely retrain the agent to behave in the way they like.

Autonomy	Low	Due to its algorithms, Amanda is not capable of making autonomous choices based on its self-determined objectives.
	High	Due to its algorithms, Amanda is capable of making completely autonomous choices based on its self-determined objectives.

Figure 11. Vignette for Manipulating Creator-Based Steerability, User-Based Steerability, and Autonomy

Measures

We measured trust as willingness to rely on the agent based on perceptions of its trustworthiness [59,87]. We adapted existing measures of trust with minimal changes to reflect the context of our study. More specifically, using a 7-point Likert scale, we adapted three items of trust used by Srivastava & Chandra [87] ("I trust Amanda to be reliable," "I believe Amanda to be trustworthy," and "I trust Amanda"). We used the same post-test questionnaire used in our previous experiments to measure our manipulations' effectiveness and several control variables.

Results

The manipulation checks confirmed that our manipulations were successful. Table 6 shows the details of Cronbach's α of the measurements, the variables' values under the low and high experimental conditions, the difference between low and high conditions, and whether the differences were significant.

Table 6. Manipulation Checks				
Variable	α	Low	High	Δ
Trust in AI Creator	0.945 (3 items)	1.702 (0.871)	4.737 (1.657)	$t(570) = 27.254, p < 0.0001, d = 2.280$; success: ✓
Creator-Based Steerability	0.959 (3 items)	2.744 (1.772)	5.940 (1.133)	$t(570) = 25.533, p < 0.0001, d = 2.136$; success: ✓
User-Based Steerability	0.990 (3 items)	2.108 (1.562)	5.947 (1.172)	$t(570) = 32.811, p < 0.0001, d = 2.744$; success: ✓
Autonomy	0.953 (5 items)	2.220 (1.299)	4.265 (1.726)	$t(570) = 16.056, p < 0.0001, d = 1.343$; success: ✓
α : Cronbach's α Δ : difference between the "low" and "high" experimental conditions (manipulation check was calculated as the average of manipulation questions) d: Cohen's d Low and High: mean values under low and high experimental conditions (standard deviations in parentheses)				

We conducted a set of hierarchical regressions to estimate the effects of trust in AI creator (low = 0, high = 1), creator-based steerability (low = 0, high = 1), user-based steerability (low = 0, high = 1), and autonomy (low = 0, high = 1) on trust in the AI agent ($\alpha = 0.981$; 3 items), while including our control variables. Table 7 shows the results.

Table 7. Regression Results for Experiment 4			
	Model 1	Model 2	Model 3
Observations	572	572	572
R ²	2.73%	22.90%	26.40%
Main Effects			
Trust in AI Creator		1.370**** (0.132)	1.656**** (0.259)
Creator-Based Steerability		-0.409*** (0.135)	-0.825**** (0.188)
User-Based Steerability		0.568**** (0.134)	1.046**** (0.187)
Autonomy		-0.314*** (0.133)	-0.114 (0.186)
Interactions			
Trust in AI Creator \times Creator-Based Steerability			0.799*** (0.262)
Trust in AI Creator \times User-Based Steerability			-0.941**** (0.261)

Trust in AI Creator × Autonomy			-0.432** (0.261)
Control Variables			
Constant	3.829**** (0.366)	3.193**** (0.361)	3.013**** (0.373)
Age	-0.001 (0.007)	-0.004 (0.006)	-0.004 (0.006)
Sex (Female=0)			
Male	0.007 (0.155)	0.014 (0.139)	0.058 (0.136)
Other	0.670 (1.257)	0.683 (1.123)	0.162 (1.106)
Education (High School or Less = 0)			
Some College	-0.087 (0.251)	-0.002 (0.225)	0.012 (0.221)
2-year College Degree	-0.215 (0.282)	-0.113 (0.253)	-0.072 (0.248)
4-year College Degree	-0.055 (0.228)	0.014 (0.204)	0.037 (0.200)
Master's Degree	0.028 (0.334)	0.009 (0.300)	0.024 (0.294)
Doctorate Degree	0.189 (0.626)	0.725 (0.563)	0.792 (0.552)
Past Experience Frequency (At least once a day=0)			
At least once a week	-0.308 (0.249)	-0.223 (0.223)	-0.215 (0.219)
At least once a month	0.211 (0.208)	0.291 (0.186)	0.253 (0.182)
Never	0.500** (0.200)	0.615*** (0.181)	0.568*** (0.178)
a. **** p<0.001, *** p<0.01, ** p<0.05, * p<0.1			
b. One-tailed tests were used for directional hypotheses			
c. Standard errors in parentheses			

Main Effects. We found support for Hypothesis 1, which predicted that trust in creator increases trust in an AI agent ($\beta = 1.370$; $p = 0.001$). Our results also provided evidence supporting the negative effect of creator-based steerability (Hypothesis 2, $\beta = -0.409$; $p = 0.001$), the positive effect of user-based steerability (Hypothesis 4, $\beta = 0.568$; $p = 0.001$), and the negative effect of autonomy (Hypothesis 6, $\beta = -0.314$; $p = 0.009$) on trust in an AI agent.

Moderation Effects. In hypothesis 3, we posited that the effect of trust in creator on trust in AI is stronger when the user perceives high creator-based steerability than when they perceive low creator-based steerability. The results provide support for this hypothesis by showing a significant positive effect ($\beta = 0.799$; $p = 0.001$). In hypothesis 5, we stated that the effect of trust in creator on trust in AI is weaker when the user perceives high user-based steerability than when they perceive low user-based steerability. This hypothesis is supported ($\beta = -0.941$; $p < 0.001$). Hypothesis 7 stated that the effect of trust in creator on trust in AI agent is weaker when the user perceives a high autonomy than when they perceive a low autonomy for the AI agent. The results support this hypothesis by indicating a significant negative effect ($\beta = -0.432$; $p = 0.049$). Figure 12 shows the interaction plots, which are consistent with those obtained in experiments 1-3.

To further confirm the robustness of our findings, we also estimated a set of hierarchical regressions that included all possible two-way, three-way, and four-way interactions (step-by-

step). Other than the hypothesized paths, however, we did not find any significant interactions ($\beta_{T \times C \times U \times A} = -0.440, p = 0.559$; $\beta_{T \times C \times U} = 0.726, p = 0.169$; $\beta_{T \times C \times A} = -0.131, p = 0.803$; $\beta_{T \times U \times A} = 0.416, p = 0.429$; $\beta_{C \times U \times A} = -0.715, p = 0.175$; $\beta_{C \times U} = 0.397, p = 0.130$; $\beta_{C \times A} = 0.206, p = 0.431$; $\beta_{U \times A} = -0.099, p = 0.705$; where T is trust in creator, C is creator-based steerability, U is user-based steerability, and A is autonomy condition). This provides further evidence that our results in experiments 1 to 3 are robust. In addition, we re-estimated the results of experiment 1 to 4 using both robust standard errors and 5,000 bootstrapping samples and found that any issues related to potential heteroscedasticity and non-normality of residuals influenced neither the direction nor the significance of our findings (see Online Appendix).

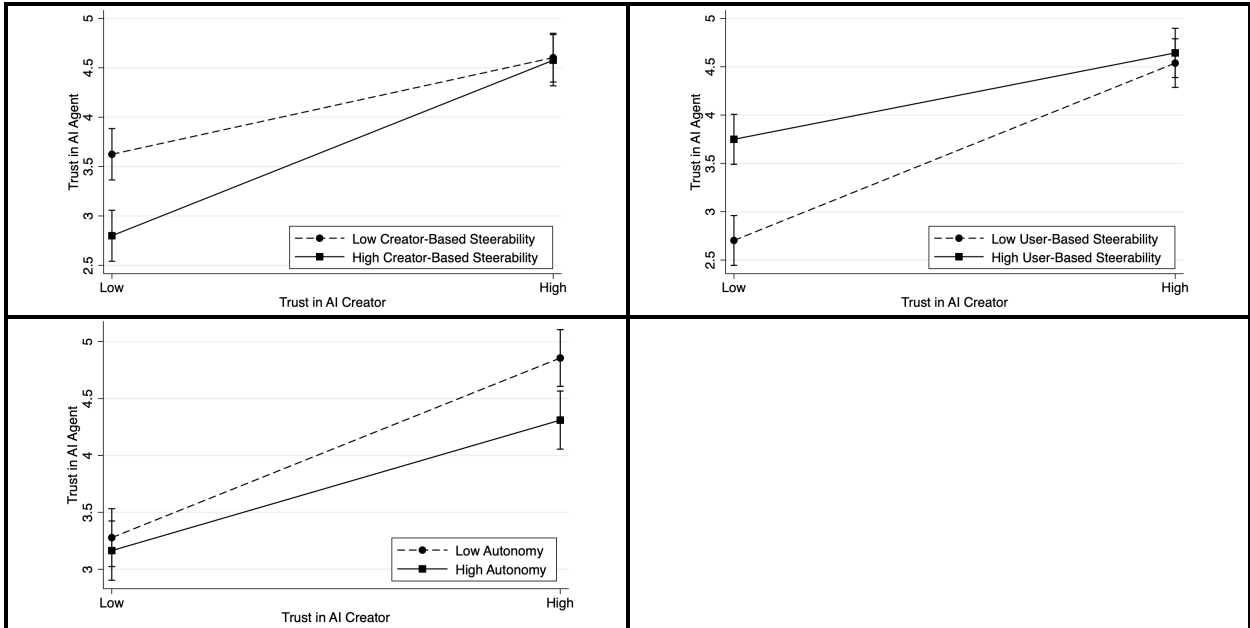


Figure 12. Moderating Role of Creator-Based Steerability, User-Based Steerability, and Autonomy

DISCUSSION AND IMPLICATIONS

In line with trust transference research [89], our experiments consistently indicate that trust in AI creator increases users' trust in AI agents (Research Question 1). However, we found that AI alignment plays an important role in this trust transference (Research Question 2). Specifically, we found that creator-based steerability, user-based steerability, and autonomy moderate trust transference, strengthening or weakening the relationship between trust in AI creator and trust in the AI agent. Comparing the results across experiments reveals interesting patterns. Particularly,

our results suggest that when people have low trust in the AI creator, they may favor an AI agent with high user-based steerability (as in Experiment 2 and Experiment 4) and with low creator-based steerability (as in Experiment 4).

Our results also provided some evidence that people are more likely to trust an AI agent with high user-based steerability, low creator-based steerability, and low autonomy.²⁰ Table 8 provides a summary of our hypotheses testing results. Before turning to the implications of our study, it is appropriate to discuss limitations.

Table 8. Summary of Hypotheses Testing Results				
Hypothesis	Exp 1	Exp 2	Exp 3	Exp 4
H1: Trust in AI creator increases trust in an AI agent.	Supported	Supported	Supported	Supported
H2: Creator-based steerability decreases trust in an AI agent.	Not Supported	-	-	Supported
H3: The positive effect of trust in AI creator on trust in an AI agent is stronger for an AI agent with high creator-based steerability than an AI agent with low creator-based steerability.	Supported	-	-	Supported
H4: User-based steerability increases trust in an AI agent.	-	Partially Supported	-	Supported
H5: The positive effect of trust in AI creator on trust in an AI agent is weaker for an AI agent with high user-based steerability than an AI agent with low user-based steerability.	-	Supported	-	Supported
H6: Autonomy decreases trust in an AI agent.	-	-	Not Supported	Supported
H7: The positive effect of trust in AI creator on trust in an AI agent is weaker for an AI agent with a high degree of autonomy than an AI agent with a low degree of autonomy.	-	-	Supported	Supported
“-” (dash) represents “not tested”				

Limitations

All research has limitations and ours is no exception. First, we explored trust transference at the initial stage of interaction with an AI agent. We acknowledge that the dynamic of trust transference might be different in later stages of the user-AI agent relationship. Second, we focused on the transfer of trust between an AI creator and the AI agent. It is possible that brand loyalty plays a role in this process. While brand loyalty itself is influenced by cognitive and affective aspects of trust [18], future research can assess whether the overall concept of brand loyalty affects users’ trust in the AI agent. Third, we identified three AI attributes by applying the agentic IS framework to the AI alignment problem and studied their effects on trust

²⁰ The evidence supporting H2 and H6 is mixed, as they were confirmed in experiment 4 but not in experiments 1 and 3. We speculate that this discrepancy may be due to differences in the experimental context. Specifically, in experiments 1 to 3, participants engaged with an interactive AI agent in the context of a hiring task, while in experiment 4 they evaluated a hypothetical AI agent in a scenario-based experiment. It is possible that differences in how people process concrete versus abstract ideas could explain the mixed results. For instance, people may find the abstract concept of an autonomous AI agent frightening [90], but may not experience the same emotions when interacting with an apparently autonomous AI agent.

transference. While our systematic approach yielded a coherent set of alignment-related AI attributes, we acknowledge that other factors (e.g., factors related to the AI agent’s alignment with other stakeholders such as the government and political parties) may also influence trust transference [2]. Further research is needed to extend our model to include such factors.

Implications for Research

In this research, we make several contributions to the literature. First, we contribute to the emerging body of literature on AI alignment by addressing the “who” question in AI alignment and proposing a stakeholder-centric approach to AI alignment. The previous literature has predominantly focused on the “what” and “how” questions, such as determining what goals and values should be embedded in AI agents [30] and how AI agents can be technically aligned to these goals and values [45,70]. Our research adds to this discourse by highlighting that different stakeholders may have different and often conflicting goals and values. As a result, AI alignment efforts should consider the entity with whom the AI goals and values should be aligned. This approach pushes AI alignment beyond the conventional aim of aligning AI with broadly conceived human goals and values [92,104], suggesting that a more nuanced, stakeholder-centric approach is needed. This shift underscores the multifaceted nature of AI alignment and the challenges that extend beyond technical solutions to encompass trust dynamics among diverse interest groups including AI creators and users [30,31].

Leveraging the agentic IS framework [2], we identified three alignment-related attributes of AI agents: creator-based steerability, user-based steerability, and autonomy. We theorized that while creator-based and user-based steerability both increase AI alignment, they have markedly different implications for trust transference due to their targets of alignment, namely, the creator versus the user. Our empirical results showed that the former positively and the latter negatively moderates trust transference, providing evidence for the importance of a stakeholder-centric approach in AI alignment efforts. We therefore propose that future research on AI alignment should consider that there are multiple stakeholders and that there may be important trade-offs

associated with optimizing for creator-AI agent alignment and optimizing for user-AI agent alignment.

Second, we contribute to the literature on human-AI interaction by highlighting the importance of a triadic view that includes the user, the AI agent, and the creator of the agent. Previous literature has primarily viewed human-AI interaction as a dyadic relationship between the user and the AI agent, ignoring the relationships between the user and the AI creator, as well as between the AI creator and the AI agent. In contrast, our research theorizes from the perspective of a triadic view, which includes all three parties. This approach allowed us to unpack the concept of AI alignment [30,36] into creator-AI alignment, user-AI alignment, and AI self-alignment, identify key constructs such as creator-based steerability, user-based steerability, and autonomy, and lay the groundwork for a more nuanced conceptual and empirical examination of AI alignment. Our triadic perspective reveals that human-AI interaction is more complex than previously understood and should be viewed within the context of the AI agent's relationship with multiple relevant stakeholders. The relationship between stakeholders with whom the AI agent aligns affects the user's trust in the agent, as well as trust transference from AI creator to AI agent. Our study provides a lens through which to understand this complex relationship. By considering the triadic view, future research can gain a more comprehensive understanding of human-AI interaction, which is essential for developing trustworthy AI agents that align with human goals and values.

Finally, we contribute to trust transference theory by identifying and exploring the boundary conditions under which trust transference breaks down or holds in the novel context of AI agents. Unlike previous research that primarily focuses on trust transfer in the context of traditional IS artifacts [82,89], our theory sheds light on the relationship between AI agents and their users. This exploration is especially important as AI agents possess varying degrees of autonomy and steerability [see 6], which raises new questions about the dynamic between the AI creator, the technology, and its users. Our contextualized theory of trust transference [40] highlights how the attributes of AI agents can impact the relationships between these entities (AI

creator, AI agent, and user), potentially disrupting the transference mechanism. Specifically, our theorization suggests that user-based steerability and autonomy can both shift a user's perception of the AI creator-AI agent relationship from a cohesive dyad to two independent entities, thereby weakening trust transference. In contrast, creator-based steerability can enhance a perception of a cohesive dyad and strengthen trust transference. Our empirical evidence supports this theory, indicating that trust transference either completely broke down (Experiment 1 to 3) or was significantly weakened (Experiment 4) when the AI agent was perceived as having low creator-based steerability, high user-based steerability, or high autonomy. Based on our theorizing and findings, future IS scholarship on trust transference should explicitly account for creator-based steerability, user-based steerability, and autonomy to gain insight into trust formation in the context of AI agents. This insight is essential for the development of trustworthy AI agents that can adapt to different contexts while maintaining alignment with human goals and values.

Implications for Practice

Practitioners can benefit from the results of this research in several ways. First, based on our study, negative news about the AI creator can not only decrease users' trust in the AI creator but can also decrease trust in the AI agent. Our moderation analyses suggest that to mitigate this transference, developers of AI agents can take three actions: (a) reduce perceived creator-based steerability, (b) enhanced perceived user-based steerability, and (c) increase perceived autonomy. Such actions can position AI agents as less reliant on their creators, insulating them from potential trust erosion resulting from a drop in their creator's reputation. As one example, to enhance perceived user-based steerability and avoid potential trust erosion, developers of AI agents LLMs can provide users with documentation or guidelines on system prompts [70]. For instance, they can inform users with liberal or conservative views that they can prompt the agent to "answer questions from a liberal [or conservative] perspective." Thus, our findings can help guide developers in designing AI agents that users will continue to trust and use over time.

With respect to increasing perceived autonomy, developers should proceed with caution as there may be trade-offs between creating a trustworthy AI agent and one that is perceived to

be independent of its creator. Specifically, while increasing perceived autonomy may alleviate problems associated with trust transference when trust in AI creator is low, our results in Experiment 4 suggest that it can decrease trust in the agent at the same time.

Finally, our findings underscore the importance of targeted AI alignment. We recommend that developers of AI agents move beyond the prevailing ‘one-size-fits-all’ alignment strategy. Given the heterogeneous goals and values among individuals, it is essential to tailor AI alignment to the specific needs of targeted users. Yet, developers must carefully navigate the technical and normative challenges of AI alignment. Creating AI agents that align with user goals and values without contravening local and international regulations is imperative.

Conclusion

With the rise of AI agents, establishing and maintaining user trust is essential. Using the agentic IS framework and trust transference theory, this study investigates the relationship between trust in AI creators and trust in AI agents, and the impact of creator-based steerability, user-based steerability, and autonomy on this relationship. We discover that while creator-based steerability boosts trust transference, user steerability and autonomy diminish it. Our findings underscore the importance of aligning AI agents’ goals and values with the appropriate entity, emphasizing a research approach that integrates the user, the AI agent, and its creator into a triadic perspective. We hope that our findings will open new doors for theory-driven research in the increasingly important area of AI agents.

REFERENCES

1. Aguirre, A., Reiner, P.B., Surden, H., and Dempsey, G. AI Loyalty by Design: A Framework for the Governance of AI. In J.B. Bullock, Y.-C. Chen, J. Himmelreich, et al., eds., *The Oxford Handbook of AI Governance*. Oxford University Press, 2022.
2. Baird, A. and Maruping, L.M. The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, 45, 1b (2021), 315–341.
3. Bakker, M., Chadwick, M., Sheahan, H., et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35, (2022), 38176–38189.
4. Bazerman, M.H. and Moore, D.A. *Judgment in Managerial Decision Making*. John Wiley & Sons, 2013.

5. Belanche, D., Casaló, L.V., Flavián, C., and Schepers, J. Trust transfer in the continued usage of public e-services. *Information & Management*, 51, 6 (2014), 627–640.
6. Berente, N., Gu, B., Recker, J., and Santhanam, R. Managing artificial intelligence. *MIS Quarterly*, 45, 3 (2021), 1433–1450.
7. Berkowitz, L. and Donnerstein, E. External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, 37, 3 (1982), 245–257.
8. Bigman, Y.E. and Gray, K. People are averse to machines making moral decisions. *Cognition*, 181, (2018), 21–34.
9. Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A., and Gray, K. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, 152, 1 (2022), 4–27.
10. Birnbaum, E. AI “Trustworthiness” Is Focal Point of New US Government Inquiry. *Bloomberg.com*, 2023. <https://www.bloomberg.com/news/articles/2023-04-11/ai-trustworthiness-is-focal-point-of-new-us-government-inquiry>.
11. Bommasani, R., Hudson, D.A., Adeli, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, (2021).
12. Broadbent, E., Kuo, I.H., Lee, Y.I., et al. Attitudes and reactions to a healthcare robot. *Telemedicine and e-Health*, 16, 5 (2010), 608–613.
13. van den Broek, E., Sergeeva, A., and Huysman, M. When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS Quarterly*, 45, 3b (2021), 1557–1580.
14. Built In. How AI Trading Technology Is Making Stock Market Investors Smarter. *Built In*, 2021. <https://builtin.com/artificial-intelligence/ai-trading-stock-market-tech>.
15. Burrell, J. How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3, 1 (2016), 1–12.
16. Burton, J.W., Stein, M.-K., and Jensen, T.B. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33, 2 (2019), 220–239.
17. Castelo, N., Bos, M.W., and Lehmann, D.R. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56, 5 (2019), 809–825.
18. Chaudhuri, A. and Holbrook, M.B. The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. *Journal of Marketing*, 65, 2 (2001), 81–93.
19. Cunneen, M., Mullins, M., Murphy, F., Shannon, D., Furxhi, I., and Ryan, C. Autonomous vehicles and avoiding the trolley (dilemma): vehicle perception, classification, and the challenges of framing decision ethics. *Cybernetics and Systems*, 51, 1 (2020), 59–80.
20. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., and Graepel, T. Cooperative AI: machines must learn to find common ground. *Nature*, 593, 7857 (2021), 33–36.
21. Das, T.K. and Teng, B.-S. Between trust and control: Developing confidence in partner cooperation in alliances. *Academy of Management Review*, 23, 3 (1998), 491–512.
22. Dasgupta, N., Banaji, M.R., and Abelson, R.P. Group entitativity and group perception: Associations between physical features and psychological judgment. *Journal of personality and social psychology*, 77, 5 (1999), 991–1003.
23. Demetis, D. and Lee, A.S. When humans using the IT artifact becomes IT using the human artifact. *Journal of the Association for Information Systems*, 19, 10 (2018), 929–952.
24. Dietvorst, B.J. and Bharti, S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, 31, 10 (2020), 1302–1314.

25. Dietvorst, B.J., Simmons, J.P., and Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144, 1 (2015), 114–126.
26. doteveryone. People, Power and Technology: The 2020 Digital Attitudes Report. 2020. <https://doteveryone.org.uk/report/peoplepowertech2020/>.
27. Fang, Y., Qureshi, I., Sun, H., McCole, P., Ramsey, E., and Lim, K.H. Trust, Satisfaction, and Online Repurchase Intention: The Moderating Role of Perceived Effectiveness of E-Commerce Institutional Mechanisms. *MIS Quarterly*, 38, 2 (2014), 407–428.
28. Floridi, L. and Sanders, J.W. On the morality of artificial agents. *Minds and machines*, 14, 3 (2004), 349–379.
29. Future of Life Institute. AI Principles. *Future of Life Institute*, 2017. <https://futureoflife.org/open-letter/ai-principles/>.
30. Gabriel, I. Artificial intelligence, values, and alignment. *Minds and machines*, 30, 3 (2020), 411–437.
31. Gabriel, I. and Ghazavi, V. The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. In C. Véliz, ed., *The Oxford Handbook of Digital Ethics*. Oxford University Press, 2022, pp. 336–355.
32. Gefen, D., Karahanna, E., and Straub, D.W. Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, 27, 1 (2003), 51–90.
33. Glikson, E. and Woolley, A.W. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14, 2 (2020), 627–660.
34. Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.
35. Gunaratne, J., Zalmanson, L., and Nov, O. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, 35, 4 (2018), 1092–1120.
36. Hadfield-Menell, D. and Hadfield, G.K. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 417–422.
37. Hagendorff, T. and Fabi, S. Methodological reflections for AI alignment research using human feedback. *arXiv preprint arXiv:2301.06859*, (2022).
38. Heider, F. *The psychology of interpersonal relations*. Psychology Press, 1958.
39. Hoffman, B.J., Bynum, B.H., Piccolo, R.F., and Sutton, A.W. Person-organization value congruence: How transformational leaders influence work group effectiveness. *Academy of Management Journal*, 54, 4 (2011), 779–796.
40. Hong, W., Chan, F.K., Thong, J.Y., Chasalow, L.C., and Dhillon, G. A framework and guidelines for context-specific theorizing in information systems research. *Information Systems Research*, 25, 1 (2014), 111–136.
41. Husch, B. and Teigen, A. Regulating Autonomous Vehicles. *National Conference of State Legislatures*, 25, (2017), 13.
42. Jussupow, E., Benbasat, I., and Heinzl, A. Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion. In *ECIS*. 2020.
43. Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence. *Information Systems Research*, 32, 3 (2021), 713–735.
44. Kee, H.W. and Knox, R.E. Conceptual and methodological considerations in the study of trust and suspicion. *Journal of Conflict Resolution*, 14, 3 (1970), 357–366.

45. Knox, W.B., Allievi, A., Banzhaf, H., Schmitt, F., and Stone, P. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316, (2023), 103829.
46. Komiak, S.Y. and Benbasat, I. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, 30, 4 (2006), 941–960.
47. Langley, P., Meadows, B., Sridharan, M., and Choi, D. Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*. 2017.
48. Lankton, N.K., McKnight, D.H., and Tripp, J. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16, 10 (2015), 880–918.
49. Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. Is AI ground truth really “true”? The dangers of training and evaluating AI tools based on experts’ know-what. *MIS Quarterly*, (2021), 1501–1525.
50. Lee, J.-G., Kim, K.J., Lee, S., and Shin, D.-H. Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction*, 31, 10 (2015), 682–691.
51. Lewicki, R.J., McAllister, D.J., and Bies, R.J. Trust and distrust: New relationships and realities. *Academy of Management Review*, 23, 3 (1998), 438–458.
52. Lewicki, R.J., Tomlinson, E.C., and Gillespie, N. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of management*, 32, 6 (2006), 991–1022.
53. Liu, G.K.-M. Perspectives on the Social Impacts of Reinforcement Learning with Human Feedback. *arXiv preprint arXiv:2303.02891*, (2023).
54. Logg, J.M., Minson, J.A., and Moore, D.A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, (2019), 90–103.
55. Longoni, C., Bonezzi, A., and Morewedge, C.K. Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46, 4 (2019), 629–650.
56. Malle, B.F. Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 18, 4 (2016), 243–256.
57. Markovski, Y. Fine-tuning a Classifier to Improve Truthfulness | OpenAI Help Center. 2023. <https://help.openai.com/en/articles/5528730-fine-tuning-a-classifier-to-improve-truthfulness>.
58. Maslej, N., Fattorini, L., Brynjolfsson, E., et al. Artificial intelligence index report 2023. *arXiv preprint arXiv:2310.03715*, (2023).
59. Mayer, R.C., Davis, J.H., and Schoorman, F.D. An integrative model of organizational trust. *Academy of Management Review*, 20, 3 (1995), 709–734.
60. McKnight, D.H., Carter, M., Thatcher, J.B., and Clay, P.F. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, 2, 2 (2011), 1–25.
61. McKnight, D.H., Liu, P., and Pentland, B.T. Trust change in information technology products. *Journal of Management Information Systems*, 37, 4 (2020), 1015–1046.
62. Microsoft. Learning from Tay’s introduction. *The Official Microsoft Blog*, 2016. <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>.
63. Microsoft. What is Azure OpenAI Service? - Azure Cognitive Services. 2023. <https://learn.microsoft.com/en-us/azure/cognitive-services/openai/overview>.
64. Mitchell, A. ChatGPT’s bias allows hate speech toward GOP, men: report. 2023. <https://nypost.com/2023/03/14/chatgpts-bias-allows-hate-speech-toward-gop-men-report/>.

65. Mullen, B. Group composition, salience, and cognitive representations: The phenomenology of being in a group. *Journal of Experimental Social Psychology*, 27, 4 (1991), 297–323.
66. Ng, A. Agentic Design Patterns. *deeplearning.ai*, 2024. <https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/>.
67. Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, (2022).
68. Olsson, A., Ebert, J.P., Banaji, M.R., and Phelps, E.A. The role of social groups in the persistence of learned fear. *Science*, 309, 5735 (2005), 785–787.
69. OpenAI. OpenAI API. 2023. <https://platform.openai.com>.
70. OpenAI. GPT-4 Technical Report. 2023. <http://arxiv.org/abs/2303.08774>.
71. Ouyang, L., Wu, J., Jiang, X., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, (2022), 27730–27744.
72. Park, J.S., O’Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., and Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, (2023).
73. Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.
74. Pavlou, P.A. and Dimoka, A. The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation. *Information Systems Research*, 17, 4 (December 2006), 392–414.
75. Ponnusamy, P., Ghias, A.R., Guo, C., and Sarikaya, R. Feedback-based self-learning in large-scale conversational ai agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, pp. 13180–13187.
76. Qiu, L. and Benbasat, I. Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems*, 25, 4 (2009), 145–182.
77. Rai, A. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, 48, 1 (2020), 137–141.
78. Robert, L.P., Denis, A.R., and Hung, Y.-T.C. Individual Swift Trust and Knowledge-Based Trust in Face-to-Face and Virtual Team Members. *Journal of Management Information Systems*, 26, 2 (September 2009), 241–279.
79. Roccas, S., Sagiv, L., Schwartz, S.H., and Knafo, A. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28, 6 (2002), 789–801.
80. Russell, S. *Human compatible: Artificial intelligence and the problem of control*. Viking, New York, NY, 2019.
81. Saharia, C., Chan, W., Saxena, S., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35, (2022), 36479–36494.
82. Schuetz, S. and Venkatesh, V. Research Perspectives: The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction. *Journal of the Association for Information Systems*, 21, 2 (2020), 460–482.
83. Schul, Y. and Peri, N. Influences of distrust (and trust) on decision making. *Social Cognition*, 33, 5 (2015), 414–435.
84. Shadish, W.R., Cook, T.D., and Campbell, D.T. Experiments and generalized causal inference. *Experimental and quasi-experimental designs for generalized causal inference*, (2002).

85. Singh, J. and Sirdeshmukh, D. Agency and trust mechanisms in consumer satisfaction and loyalty judgments. *Journal of the Academy of Marketing Science*, 28, 1 (2000), 150–167.
86. Smith, H.J., Dinev, T., and Xu, H. Information privacy research: an interdisciplinary review. *MIS Quarterly*, 35, 4 (2011), 989–1016.
87. Srivastava, S.C. and Chandra, S. Social Presence in Virtual World Collaboration: An Uncertainty Reduction Perspective Using a Mixed Methods Approach. *MIS Quarterly*, 42, 3 (2018), 779–803.
88. Stewart, K.J. Trust transfer on the world wide web. *Organization Science*, 14, 1 (2003), 5–17.
89. Stewart, K.J. How hypertext links influence consumer perceptions to build and degrade trust online. *Journal of Management Information Systems*, 23, 1 (2006), 183–210.
90. Szollosy, M. Freud, Frankenstein and our fear of robots: projection in our cultural perception of technology. *AI & SOCIETY*, 32, 3 (2017), 433–439.
91. The Economist. Is Google’s Gemini chatbot woke by accident, or by design? 2024. <https://www.economist.com/united-states/2024/02/28/is-googles-gemini-chatbot-woke-by-accident-or-design>.
92. United Nations. *Our Common Agenda – Report of the Secretary-General*. United Nations, New York, NY 10017, 2021.
93. Vincent, J. OpenAI CEO Sam Altman on GPT-4: “people are begging to be disappointed and they will be.” *The Verge*, 2023. <https://www.theverge.com/23560328/openai-gpt-4-rumor-release-date-sam-altman-interview>.
94. Wang, N., Shen, X.-L., and Sun, Y. Transition of electronic word-of-mouth services from web to mobile context: A trust transfer perspective. *Decision support systems*, 54, 3 (2013), 1394–1403.
95. Wang, W. and Benbasat, I. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23, 4 (2007), 217–246.
96. Wang, W. and Benbasat, I. Attributions of trust in decision support technologies: A study of recommendation agents for e-commerce. *Journal of Management Information Systems*, 24, 4 (2008), 249–273.
97. Wang, W. and Benbasat, I. Research note—A contingency approach to investigating the effects of user-system interaction modes of online decision aids. *Information Systems Research*, 24, 3 (2013), 861–876.
98. Wang, W., Xu, J., and Wang, M. Effects of Recommendation Neutrality and Sponsorship Disclosure on Trust vs. Distrust in Online Recommendation Agents: Moderating Role of Explanations for Organic Recommendations. *Management Science*, 64, 11 (2018), 5198–5219.
99. Wang, Y., Yao, Q., Kwok, J.T., and Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53, 3 (2020), 1–34.
100. Woodside, A.G. and Chebat, J.-C. Updating Heider’s balance theory in consumer behavior: A Jewish couple buys a German car and additional buying–consuming transformation stories. *Psychology & Marketing*, 18, 5 (2001), 475–495.
101. Xu, J., Benbasat, I., and Cenfetelli, R.T. The nature and consequences of trade-off transparency in the context of recommendation agents. *MIS Quarterly*, 38, 2 (2014), 379–406.

102. Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32, 4 (2019), 403–414.
103. You, S., Yang, C.L., and Li, X. Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems*, 39, 2 (2022), 336–365.
104. Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M.C., and Dafoe, A. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research*, 71, (2021), 591-666-591–666.