The Business School for the World®

Working Paper

2023/47/STR

(Revised version of 2021/42/STR)

Artificial Intelligence, Trust, and Perceptions of Agency

Bart S. Vanneste
University College London, b.vanneste@ucl.ac.uk

Phanish Puranam
INSEAD, Phanish.puranam@insead.edu

25 August 2023

forthcoming in *Academy of Management Review*

The literature on trust among humans assumes that trustees are viewed as having agency (i.e., they display the capacity to think, plan and act), else trust is undefined. In contrast, the literature on confidence in technology does not require this assumption about the technologies we make ourselves vulnerable to (thus the term "confidence" rather than the term "trust" when applied to technology). Modern artificial intelligence (AI) technologies based on deep learning architectures are often perceived as agentic to varying degrees—typically as more agentic than other technologies but less than humans. We theorize how different levels of perceived agency of AI affect human trust in AI. We do so by investigating three mechanisms. First, a more agentic seeming AI (and its designer) will appear more able to execute relevant tasks, and therefore more trustworthy. Second, the more agentic seeming the AI, the more important are trustworthiness perceptions about the AI relative to those about its designer. Third, because of betrayal aversion, the anticipated psychological cost of the AI violating trust increases with how agentic it seems to be. These mechanisms imply, perhaps counterintuitively, that making an AI appear more agentic may increase or decrease the trust that humans place in it: success at meeting the Turing test may go hand in hand with a decrease of trust in AI.

Keywords: Artificial Intelligence; Trust; Agency; Betrayal Aversion
JEL Classification: M00, M14, Z10

Electronic copy available at: https://ssrn.com/abstract=3897704

Acknowledgement: We thank the Editor, Laura Poppo, and three anonymous reviewers for their valuable comments and suggestions throughout the review process. For comments on an earlier version of this paper, we thank Angela Aristidou, Anil Doshi, Vivianna Fang He, Tianyu He, Xi Kang, Georg von Krogh, Steven Lui, Oliver Schilke, Yash Raj Shrestha, and participants of the UCL School of Management reading group and the Rapid Research seminar at INSEAD. We are also grateful for the feedback of participants of the Idea Development Workshop for the Special Topic Forum on "Fresh Perspectives on Trust." Puranam thanks the Desmarais Fund at INSEAD for supporting the Organizations and Algorithms project.

INTRODUCTION

Artificial intelligence (AI) is seen today as a transformative technology that has created the possibility of large-scale human reliance and possibly dependence on non-human intelligences (Shrestha et al., 2019; Csaszar and Steinberger, 2021). The number of arenas in which humans interact with and rely on AI has increased dramatically (Agrawal et al., 2018; Raisch & Krakowski, 2020; Lindebaum et al., 2020), with prominent use-cases in transportation (driver-assisted or autonomously driven cars), at home (virtual assistants), in healthcare (AI-interpreted scans), in knowledge work (predictive analytics, natural language processing (NLP), and most recently through generative AI models like ChatGPT and Dall-E2). Recent work has started to investigate when and why humans trust AI (e.g., Glass et al., 2008; Waytz et al., 2014; Dietvorst et al., 2015, 2018; Verberne et al., 2015; De Visser et al., 2016; Ullman & Malle, 2018; Glikson & Woolley, 2020; Lockey et al. 2021). This work draws on two foundational literatures: one focusing on the *trust that humans place in fellow humans* (Mayer et al., 1995; Bhattacharya et al., 1998; Rousseau et al., 1998; Colquitt et al., 2007; Vanneste et al., 2014); and another concerned with human *confidence in technology* (Parasuraman & Riley, 1997; Lee & See, 2004; Hancock et al., 2011; McKnight et al., 2011; Hoff & Bashir, 2015).

These literatures however reflect fundamentally different assumptions about the perceived agency of humans and technological systems, respectively. The perception of agency is the perception that an entity has the ability to think, plan, and act (Gray et al, 2007; Gray and Wegner, 2012). The literature on trust in humans requires that the trustee (i.e., a fellow human) is seen to act with agency, else trust is undefined (Rousseau et al., 1998; Hardin, 2002). Mayer et al. (1995: 712) provide a widely accepted definition of trust in humans as “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party". They add that "[t]his definition of trust is applicable to a relationship with another identifiable party who is perceived to act and react with *volition* toward the trustor" (p. 712, emphasis added).[1]

In contrast, the literature on confidence in technology does not require that a technology must appear to have agency for people to have confidence in it (Parasuraman & Riley, 1997). For example, McKnight et al. (2011) investigated user's confidence in Microsoft Excel, a spreadsheet program. However, this distinction is not always clear in the case of AI. Modern AI technologies, which learn how to act rather than simply obey programmed instructions, occupy a curious place somewhere between humans and inanimate technology in the extent to which they are seen as agentic (Kim & Hinds, 2006; Waytz et al., 2014; Banks, 2019; Lee et al., 2019). We propose that taking into account how agentic they are perceived to be yields not only a more comprehensive account of human trust in AI but also a unique understanding of when we are more likely to trust or not trust AI.

The purpose of this paper, therefore, is to understand the role of agency perceptions about AI in human trust in AI. We specify three mechanisms through which the degree of agency perceptions can affect human trust in AI. First, if the AI is seen as more agentic, perceptions about its ability as well as that of its designer are strengthened (and therefore so are the perceptions about their trustworthiness). This mechanism involves *agency as enhancing ability perceptions*. Second, compared to the trustworthiness perceptions about its designer, the trustworthiness perceptions about the AI become more important as the AI is seen as more agentic. This mechanism involves a *shift in the locus of trust due to perceived agency*. Third,

[1] Chadderdon (2008: 254) defines volition as "the capacity for adaptive decision making", akin to the definition of agency as the ability to think, plan, and act (Gray et al, 2007; Gray and Wegner, 2012) that we use.

because of *aversion to betrayal by an agent* (see Bohnet & Zeckhauser, 2004), the anticipated psychological cost if the AI violates trust increases with how agentic it seems to be. This mechanism reduces the human's trust in AI.

Perhaps counterintuitively, these mechanisms jointly imply that making an AI appear more agentic need not increase the trust that humans place in it. Thus, while designers of modern AI technology often strive to endow their creations with features that make them appear more human (e.g., through anthropomorphizing or giving them a degree of autonomy), this may also change agency perceptions and so may make a human less keen to trust an AI. Using our conceptual model, we predict which kinds of interventions for enhancing human trust in AI are likely to succeed or fail. We also take a dynamic perspective, considering how the development of ex-post trust (i.e., after the outcome of trusting becomes known) over time is affected by agency perceptions, which may themselves change over time.

The main contribution of our paper is to extend the existing trust literature by converting agency perceptions from a boundary condition—usually assumed to be met when discussing trust among humans—to a variable. We argue that agency perceptions not only determine the relevance of the notion of trust in AI, but also influence its level. We propose that taking the role of agency perceptions into account will help us better understand the conditions under which humans trust AI, and how this is different from trust in fellow humans or confidence in technology that is not perceived as agentic. Further, we highlight mechanisms through which agency perceptions change ex-ante trust (before an interaction) as well as ex-post trust (i.e., after the outcome of trusting is known), in ways that go beyond simply increasing or decreasing relationships. Our approach helps us uncover commonalities in the diversity of previous findings (e.g., Glikson & Woolley, 2020), as well as make new falsifiable predictions for testing in future studies.

THE IMPORTANCE OF AGENCY PERCEPTIONS FOR TRUST

We begin by highlighting the role of perceived agency in discussions of trust among humans. We argue that the perception of agency is an explicit part of the classic definitions of trust (such as in Mayer et al., 1995 or Rousseau et al., 1998), but its level is implicitly assumed to be constant (and usually high) when dealing with human trustees. We relax this assumption and note that, if perceptions about the agency of a trustee can vary in strength, this will have implications for trust. Our theory thus makes the perception of agency a variable that can vary not only between human and AI as trustees, but also with types of AI as a function of their attributes. Next, we discuss the factors that underlie variations in perceptions of agency, arguing that the modern approach to AI produces systems that are much more likely to be perceived to have agency than older AI or other technologies.

Trust Requires a Perception of Agency

Mayer et al.'s (1995: 712) classic definition of trust makes clear that trust requires the trustee to be seen as having agency. If the trustee is likely to perform the action because they have no choice in the matter, there is no need to trust them. Another influential treatment of trust is from Rousseau et al. (1998), who emphasize that trust relies on expectations of positive intentions. This role of agency (because of the focus on volition and intention) in these definitions underscores the point that a trustee should be someone with perceived agency or seen as someone who is a source of "planful action" (Taylor, 1985; Morris et al., 2001). For example, playing a lottery against nature (e.g., living in a floodplain), even if risky, is not an instance of trust. Participating in a clinical trial in the hope of accessing a last resort treatment, where access to the life-saving treatment is randomly assigned, is also not a matter of trust in the experimenter, who is known to be randomizing rather than helping specific people. Trust is relevant only because the trustee has some agency, and therefore a choice between acting in a way that may harm or benefit the trustor. Hardin (2002: 12) captures well the centrality of agency attributions to trust:

> *"For example, consider an extreme case: I am confident that you will do what I want only because a gun is pointed at your head. (I have grasped the wisdom of Al Capone, who is supposed to have said, "You can get so much farther with a kind word and a gun than with a kind word alone" [quoted in McKean 1975, 42n].) Part of what is wrong when I coerce you to do what I "trust" you to do is that such an act violates the sense that trust as a concept has no meaning in a fully deterministic setting. I do not trust the sun to rise each day, and if people were fully programmed robots I would not in our usual sense trust them."*

Here Hardin makes the case that there is no trust when the trustee is perceived to have no agency. Such a binary distinction makes sense when contrasting humans with most other systems that we interact with or depend upon, but not all. In particular, we propose that AI technologies evoke varying degrees of agency perception in the minds of their human observers, with important implications for human trust in AI.

Perceptions of Agency Can Vary across Technologies

Researchers have examined the perceptions of agency that humans hold about non-human entities in terms of these entities' abilities to think, plan, and act (Gray et al, 2007; Gray and Wegner, 2012). For instance, Banks (2019) developed a scale of perceived agency by showing participants short video clips of a chatbot, voice assistant, on-screen agent, robot, or a human.

Whereas the artificial agents were rated as less agentic than the human, participants were comfortable using the same scale for both artificial and natural agents. In interacting with technological systems, humans thus perceive different degrees of agency in the system based on its attributes. For instance, Kim & Hinds (2006) found in a study of robots used in hospitals that when a robot is more autonomous (i.e., can exercise control over its own actions), people attribute more credit and blame to the robot and less to themselves and other participants. Waytz et al. (2014) found that humans' trust in self-driving cars increases with the extent of their anthropomorphization, because this makes them appear to have a mind. Lee et al. (2019) found that people adjusted their negotiation approach with virtual agents depending on the communication style and the perceived agency of the virtual agent.

A conceptual and empirical literature has identified two preconditions for agency perceptions: perceptions of intentionality and free will (e.g., Premack & Woodruff, 1978; Clarke, 1995; O'Connor, 1995; Kim & Hinds, 2006; Gray et al., 2007; Baumeister & Monroe, 2014, Waytz et al., 2014). Intentionality in this literature refers to the capacity to take goal-oriented decisions based on a logic of how to accomplish those goals under changing circumstances (Bigman et al, 2017; Van der Woerdt & Haselager, 2019). It involves understanding how actions lead to outcomes. Perceptions of free will involve attributing the capacity to select one's own goals autonomously (Gray and Wegner, 2012).[2] It involves being perceived as able "to do otherwise". These perceptions that an entity can understand an action's consequences

[2] The concept of free will has been important across a number of philosophical traditions and continues to engage much scholarly interest among both philosophers and social scientists. As can be expected for any topic that has attracted more than two millennia of interest and still seems to have many unresolved aspects, we will be unable to provide a complete account of past thinking on free will (readers are referred to the useful entries on the topic in the Stanford Encyclopedia of Philosophy [https://plato.stanford.edu/entries/freewill/]). Instead, we focus on the perception of free will which has been argued to be an important component of the perception of agency (e.g. Gray et al., 2007; Gray and Wegner, 2012).

(“intentionality”), and is capable of selecting actions freely (“free will”) contribute to producing agency perceptions, i.e., the belief an entity can think, plan, and act.

Modern AI Creates Stronger Agency Perceptions

An ambitious goal for computer scientists has been to create artificial intelligence that mimics or exceeds natural human intelligence (Samuel, 1959). Intelligence is understood as “an agent’s ability to achieve goals in a wide range of environments” (Legg and Hutter, 2007: 402). The capacity to multiply big numbers at high speed is not considered sufficient to attain this goal; it is not even necessary, as people who fall far short of the astonishing computational speeds achieved by computers still show behavior that is classified as intelligent. Instead, intelligence is associated more generally with reasoning, learning, decision making, problem solving, and other higher-order thinking skills (Gottfredson, 1997).

The OECD (2022) defines an AI system as “a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.” This definition reflects the broad scope and diversity among AI systems. For our theory, two attributes of AI systems are relevant. First, they create agency perceptions among humans interacting with them, which vary according to the system’s degree of perceived autonomy, anthropomorphy, and opacity (Kim & Hinds, 2006; Chadderdon, 2008; Bigman et al., 2019). Second, they create potential vulnerabilities for the user, which make trust relevant. Trust is more important in high-vulnerability contexts. The degree of vulnerability depends on the task and ranges from low (as in a movie recommendation) to high (as in self-driving cars or medical diagnoses).

We argue that modern AI systems are likely to be seen as more agentic because they rely on the connectionist approach, the dominant approach in AI today (Bengio et al., 2021). Connectionism is named for the connections between nodes in neural networks. It is based on work in cognitive science that views the human brain as a collection of basic neurons that make possible the performance of complicated logical calculus through their learned connections (McCulloch and Pitts, 1943). This approach focuses on algorithms learning (or "machine learning") from data or interactions with an environment.[3] The connectionist approach stands in contrast to the symbolic approach, or Good Old-Fashioned Artificial Intelligence (GOFAI) (Haugeland, 1985), which was dominant before the 1990s. This older approach used symbols and their manipulations (e.g., rules, heuristics, and logic) pre-defined by a human designer to develop intelligent systems (Newell & Simon, 1976). For example, a chess computer based on old AI executes good chess moves based on rules provided by its human designer. In contrast, one based on modern AI learns good chess moves from data.

The connectionist approach is inherently likelier than the older rule-based approach to lead to AI behavior that is perceived as agentic for two reasons (Kim & Hinds, 2006). First, agency perceptions increase with autonomy (Chadderdon, 2008; Wyatz et al., 2014), i.e., whether a system "exercises control over its own actions" (Franklin & Graesser, 1997: 29). Autonomy is positively related not only to free will perceptions (because of the possibility for a

[3] For example, in supervised learning the data consists of inputs and an output. The goal is to learn a function that describes the relationship between these inputs and output (e.g., using CVs to predict job performance). In unsupervised learning the data has no specific output. Instead the goal is to uncover a plausible compact description of the data through clustering or dimension reduction (e.g., grouping related documents together). In reinforcement learning the data are generated through interacting with an environment, and the goal is to find a series of optimal actions to reach a goal (e.g., controlling a chemical plant, or in a self-driving car) (for an accessible overview of types of machine learning algorithms, see Shreshtha et al., 2021).

system to select its own actions), but also intentionality perceptions (because of the system's appearance of exhibiting intentions by making judgements (Kim & Hinds, 2006)).

Older symbolic (or rule-based) AI, in which decision rules are hardwired by the system's designer, is perceived to have low agency because it merely follows pre-defined rules and its own actions are constrained by design. In contrast, modern, connectionist AI differs primarily in the fact that it is capable of inducing its own arbitrarily complex rules of behavior from data, either from archival records or in real time. This implies that its behavior can be perceived as less subject to constraints imposed by a designer. Whereas old AI follows pre-defined rules, modern AI infers rules from the patterns found in data and can change its rules following new data or interactions. This difference leads to greater autonomy perceptions (Boden, 1996). Thus, greater autonomy of modern AI systems leads to them being perceived as more agentic.

The second reason that modern AI is more likely to be seen as agentic than old AI is because of the high opacity of its workings (Kim & Hinds, 2006; Lyons et al., 2023), that is, the low level of a user's understanding of why a particular decision was made (Sinha & Swearingen, 2002). Transparency diminishes perceptions of free will, whereas opacity preserves it (Bigman et al., 2019). For example, people exposed to scientific advances in neuroscience on the mechanistic causes of human decision-making attribute weaker free will to others (Shariff et al., 2014). If an AI's inner workings are clear to observers, then agency is not needed to explain its behavior. In contrast, if the reasons underlying an AI's behavior are opaque, then it is more likely seen as itself responsible for producing that behavior (Kane, 1996; Chadderdon, 2008).

Recall that old AI adopts human-made rules, whereas modern AI discovers its own rules. The success of modern AI is most visible precisely in situations where rules are too complex to describe for humans. For instance, a neural network is built up of layers through which input data is transformed into an output, e.g., a decision. Each successive layer allows for a more complex decision rule.[4] Indeed, expanding the size of the network and the number of layers produces arbitrarily complex decision rules. This is why neural networks are “universal function approximators”, capable of producing decision rules of arbitrary complexity (Hornik et al., 1989). Because the rules induced in modern AI can be far more complex and challenging to understand than what can be symbolically coded by a human programmer in old AI, modern AI can be perceived to have greater opacity and therefore a greater degree of agency.

In sum, modern AI is seen as more agentic than old AI because in modern AI decision rules are self-learnt and can change as opposed to fixed and dictated by a human programmer (autonomy) and the decision rules are typically much more complex and hard to understand for human users (lack of transparency).[5] To be sure, modern AI may still fall far short of being perceived as comparable to a human in terms of agency. When dealing with fellow humans, we typically assume significant levels of agency on their part, with exceptions made for conditions of ill health, age, or duress. Agency perceptions about modern AI are therefore likely to occupy a level somewhere between that for older AI (and in fact most technology) and fellow humans.

Consider, for instance, the difference between taking investment advice from a regression model built by experts, a connectionist AI that interacts with the user, and a human investment manager who makes investment recommendations. The regression model is likely to be perceived as having the least agency and the human, the most. The interactive AI is likely to be perceived to have a level of agency that varies between these two cases (Banks, 2019; Lee et al.,

[4] The usage of many layers is the source of the term “deep” in deep learning, which is a popular term for multi-layered neural networks.

[5] A potential downside of increasingly complex algorithms is that they may “overfit”, i.e., when they capture not only the systematic relationships between inputs and outputs but also random noise in the data that was used for building the complex decision rules. To the extent that overfitting is more likely with complex decision rules, that could introduce a link between overfitting and agency perceptions. We thank a reviewer for pointing out a connection between these.

2019; Murray et al., 2020), because it has greater perceived autonomy and is more opaque than the regression model. In addition, if the AI is embodied in a chatbot, this may increase anthropomorphy (another source of agency perception), though in principle the regression model can also be encapsulated in a chatbot. The more important difference, in our view, is that the newer AI technologies display greater autonomy and opacity, which is why they may be perceived to have greater agency than the older systems, other things being equal.

In the next section we develop our theory about the implications of varying levels of perceived agency for trust in modern AI.

A MODEL OF PERCEIVED AGENCY AND TRUST IN AI

Existing Theory

Our model of human trust in AI builds on the existing literature in two ways. First, we build on the idea that a key driver of trust is perceived trustworthiness (Mayer et al., 1995) and second on the importance given not only to the technology that humans rely on, but also its designer (Parasuraman & Riley, 1997; Hoff & Bashir, 2015).

The first input to our model is that the trust literature distinguishes trust from perceived trustworthiness (Mayer et al., 1995; Bhattacharya et al., 1998; Rousseau et al., 1998). Trust is a trustor's willingness to be vulnerable based on some positive expectations about the other. These positive expectations have been referred to as perceived trustworthiness, or the trustor's perceptions about the trustee's trustworthiness. Because trustworthiness is not directly observable, the trustor can at best form a perception about the trustee's trustworthiness (Vanneste et al., 2014).

Perceived trustworthiness rests on perceptions about ability, benevolence, and integrity (Mayer et al., 1995; Williams, 2001; Colquitt et al., 2007; Weber and Bauman, 2019). Ability focuses on the trustee's competencies and skills needed to perform the task important to the trustor. Benevolence is the extent to which the trustee acts in the best interests of the trustor, regardless of any personal gains. Integrity refers to the extent to which a trustee adheres to sound moral and ethical principles. Thus, perceived ability captures whether the trustee is believed to have "can-do" for the focal task. Perceived benevolence and integrity together capture whether the trustee is believed to have "will-do" for the focal task (Colquitt et al., 2007).

Ability, benevolence, and integrity perceptions are invoked not only about humans but also about technologies (McKnight et al. 2002; Vance et al., 2008; McKnight et al., 2011; Lankton et al., 2015). Ability perceptions about AI capture the belief that it has the skills, competencies, and expertise to function adequately and reliably. Ability perceptions have also been referred to as perceptions about competency, functionality, and reliability (McKnight et al., 2002, 2011). The rapid development of AI has led to a significant increase in expectations regarding its technical capabilities, potentially creating unrealistic expectations (Burton et al., 2020). Ability perceptions are a key determinant of trust in AI (Gulati et al., 2018) and for its adoption (Davis, 1989; Venkatesh & Davis, 2000).

AI benevolence perceptions indicate whether the AI is believed to act in the best interest of the user. While present AI is incapable of forming genuine attachments, users may nonetheless perceive such attachments (Xu et al., 2022). A common theoretical explanation is the Computer-Are-Social-Actors (CASA) paradigm (Nass & Moon, 2000; Nielsen et al., 2022), which has documented that people's interactions with computers are essentially social because they apply rules, norms, and expectations similar to those in interpersonal relationships when interacting with computers. One famous example is a study on reciprocity (Fogg and Nass, 1997), in which participants received helpful search results from a computer. In a subsequent task, half of the participants spent time helping the same computer generate color palettes. However, the other half were switched to a different computer and were found to spend less time helping. Thus, people reciprocated help from the original source, a computer, just as they would if it were a human. Comparable social attitudes have been observed toward AI such as healthcare robots, virtual assistants, self-driving cars, customer support agents, and smartphones (Wang, 2017; Edwards et al., 2019; Kim et al., 2019; Westerman et al., 2020; Pelau et al., 2021). Computers are seen as social entities based on cues and signals, even if minimal (Nass & Moon, 2000), including interactivity, engagement, autonomy, learning, and unpredictability (Nass, 2004; Fiore et al., 2013; Feine et al., 2019). These qualities are typical of connectionist AI. When AI is treated as a social actor, it can lead to perceptions of their benevolence or lack thereof (Brave et al., 2005; Lee & Nass, 2010; Xu et al., 2022).

AI integrity perceptions refer to the belief that the AI is honest, fair, and truthful (McKnight et al. 2002, Höddinghaus et al. 2021). Although current AI cannot adhere to a set of moral and ethical principles in the way humans can, it can operate according to rules that are shaped by its designer and learned from data. These rules lead to behaviors and outcomes that can be perceived as more or less fair. An active research area exists on algorithmic fairness that investigates the extent to which the algorithms underlying AI systems are seen as fair (Oneto & Chiappa, 2020; Mitchell et al., 2021; Kallus et al., 2022; Starke et al., 2022). Just as for human decision making, one can distinguish here between procedural fairness and outcome fairness (Barocas & Selbst, 2016). Procedural fairness is focused on determining whether an algorithm treats users differently based on their membership in a protected class, such as race or gender (“disparate treatment”). For example, does a financial algorithm make the same loan recommendation for two people who are identical except for their protected class membership? Outcome fairness is focused on determining whether an algorithm adversely affects users of one protected class more than another (“disparate impact”), even if the algorithm treats two otherwise identical users alike. For example, does the financial algorithm approve a higher proportion of loans for people from one class than another? A perceived lack of fairness, and therefore of integrity can lead to lower trust in AI (Woodruff et al., 2018; Shin et al., 2020).

The ABI (ability-benevolence-integrity) perceptions model of trust in humans has been successfully applied to AI. For example, users have been asked about their ability, benevolence, and integrity perceptions of Siri, Apple’s virtual assistant (Gulati et al., 2018), recommender systems (Benbasat & Wang, 2005), and human resources decision algorithms (Höddinghaus et al. 2021). Statements to measure these perceptions include “Siri is a capable and efficient intelligent personal assistant” for ability (Gulati et al., 2018), “This virtual advisor puts my interests first” for benevolence (Benbasat & Wang, 2005), and “I think [the computer program] makes unbiased decisions” for integrity (Höddinghaus et al. 2021).

The second input to our model is that the technology literature emphasizes the importance of the designer of a technology for humans to place confidence in the technology (Lee & Moray, 1992; Parasuraman & Riley, 1997). For example, when deciding on riding in an autonomous vehicle, a passenger cares not only about attributes of the vehicle but also of its designer (Waung et al., 2021). Likewise, the willingness to depend on an AI algorithm depends on the trust in the technology company that designed it (Lumineau et al., 2020; Zhang et al., 2021). More generally, because many technologies are seen as only partially responsible for outcomes, consideration is given to the technology's creator (Lee & See, 2004; Hancock et al. 2011).

Thus, people's trust in AI reflects in part their trustworthiness perceptions of its designer. One may summarize these relationships by noting that a user's relationship with a technology changes from dyadic to triadic when taking into account the designer (Hoff & Bashir, 2015). With an agentic technology, a user can have trust in the technology and/or its designer (see Figure 1). With a non-agentic technology, a user can have confidence in the technology (because we reserve the term "trust" for trustees with perceived agency), but may have trust in the designer. In this way, confidence in a non-agentic technology and trust in a human are conceptually closer than may appear apparent at first sight (Culley & Madhavan, 2013; Hoff & Bashir, 2015). In the situations we analyze, the trustor is always the user. The trustee is the AI and/or the designer, depending on the perceived agency of the AI in the eyes of the user.

[[INSERT FIGURE 1 ABOUT HERE]]

The term "designer" refers to the person(s) responsible for building the AI. This may refer to a single human inventor or a group of human inventors. However, a user may more typically think of the designer as the organization to which the human inventor(s) belong(s) (Hengstler et al., 2016).[6] An extensive literature has investigated the facilitators and impediments to technology adoption (Davis, 1989; Venkatesh et al., 2003). Not only do the characteristics of the technology play a key role but also those of the firm that produces the technology (Nienaber & Schewe, 2014). Attention has especially been paid to a user's trust in that firm or its perceived trustworthiness (Kim et al., 2008; Luo et al., 2010; Wu et al., 2011; Williams et al., 2015). For example, the acceptance of a chatbot for disease diagnosis in healthcare depends on not only trust in the system but also in the company behind the system (Laumer et al., 2019).

[6] We thank a reviewer for highlighting this organizational perspective.

To be sure, the identity of the AI's designer is not always known. This situation is more common if the designer refers only to the person(s) who built the AI, but less common if it includes the collection of people (i.e., organization) responsible for building the AI (as we mean in our usage). The perceived trustworthiness of the designer can only affect trust in the AI if the designer is known, which is thus the situation we focus on. Knowing the designer's identity differs from having previously interacted with the designer. People can and do have trustworthiness perceptions about others with whom they had limited, or even no, prior interactions (Nannestad, 2008; Vanneste & Gulati, 2022). In this case, a user may attribute trustworthiness levels to a designer based on the user's interactions with other companies (Zhang & Yencha, 2022).

Figure 2 shows our model of human trust in AI before commencing interacting with it. This ex-ante level of trust (i.e., trust before interacting) may be modified based on the outcome of interacting with the AI, which can lead to changed ex-post levels of trust. We consider this modification after interaction separately. For now, our discussion of trust is restricted to the ex-ante trust, which is based on the relationships established in the existing literature (shown with full, black arrows). First, because perceived trustworthiness of the AI positively contributes to trust in the AI, the AI's perceived ability, benevolence and integrity are positively related to trust in the AI. Second, because trust in AI also depends on the perceived trustworthiness of the designer, the designer's perceived ability, benevolence and integrity are also positively related to trust in the AI.

[[INSERT FIGURE 2 ABOUT HERE]]

Adding Perceived Agency of AI

To these established relationships, we add the role of how agentic the AI is perceived to be by the human. We extend the existing literature by highlighting three mechanisms through which these agency perceptions extend the baseline model: (1) enhancing perceived ability, (2) shifting the locus of trust, and (3) amplifying betrayal concerns.

Mechanism 1: Enhancing Agency Perceptions. The first mechanism added to the baseline model is the idea that the more an AI is perceived as agentic, the more able it is seen to be. For example, we know that more lifelike robots (that are able to move body parts or express facial emotions) or more humanlike robots (that physically resemble humans) are perceived as more agentic and as more able to discharge their duties (De Ruyter et al., 2005; Powers et al., 2007; Belanche et al., 2021).

Moving beyond human appearance, research on mind perception has linked perceptions of agency with those of ability through two mechanisms (Fiske et al., 2002; Fiske et al., 2007; Gray et al., 2007; Gray & Wegner, 2012). The first is a direct mechanism. Agency perceptions make an AI appear likelier to successfully complete a task (Waytz et al., 2014). The perceived capacity for selecting (sub-)goals (free will) and purposeful actions (intentionality) gives an AI a greater perceived range of possibilities to learn from the environment and adjust to changes in the environment, or a greater toolset to complete a task. Agency perceptions appear especially beneficial in situations that require a (re-)assessment of goals and actions. For example, a robotic surgeon may be perceived as more able to operate on a patient when it is seen as reacting and adjusting to real-time video of a patient's organs than when it is seen as following a set of predefined instructions.

The second is an indirect mechanism. Agency perceptions are both causes and effects of perceived anthropomorphism, that is, perceptions that an AI is humanlike. Thus, if a user views an AI as having agency, the user will also tend to attribute other human abilities to the AI, including ability (Krach et al., 2008; Rosenthal-Von Der Pütten & Krämer, 2014; Belanche et al., 2021; Hu et al., 2021). If the AI is seen as more similar to a human, then the expectations of successful task completion may increase. In this instance, compared to a non-agentic-seeming robotic surgeon, we may perceive an agentic-seeming robotic surgeon as more able because of its real or imagined resemblance to a human surgeon.

We highlight that though they may be correlated, perceived ability and perceived agency are distinct. Ability captures the competencies and skills necessary for completing a task (Mayer et al., 1995; Colquitt et al., 2007). We can contrast it with the two key components of agency perceptions: intentionality and free will perceptions. Recall that intentionality refers to the capacity to hold the belief that an action will lead to an outcome (e.g., Weisman et al, 2017; Bigman et al, 2019). A system can be competent ("fit for purpose") without being intentional, e.g., a type-writer. Likewise, a system can display intentionality without being competent, e.g., a self-driving car that causes frequent accidents. Similarly, because free will is the capacity to choose one's own goals, and ability is the competence to execute conditional on goals, one can exist without the other (Shariff et al., 2014; Bigman et al, 2019). For instance, an AI algorithm embedded in a robotic exploration device that can choose its own immediate goals may appear to have free will, but it may be perceived to lack ability if it fails at accomplishing those goals. Conversely, a self-driving car may competently take passengers from one location to another, but it does not have the capacity to choose its own destinations (i.e., it will not be perceived to have free will, even though it may have intentionality). Thus, although we say that an increase in perceived agency should lead to an increase in perceived ability, we do not treat these two constructs as identical.

Not only will a more agentic AI be seen as more able, but also, we argue, so will its designer. Because ability is not directly observable (Mayer et al., 1995), a user will consider actions and artifacts to make inferences about the designer (Anderson & Weitz, 1989; Vanneste et al., 2014). Designing an agentic AI is harder than designing a non-agentic AI, so more agentic perceptions of the AI should also lead to greater ability perceptions of the AI's designer. This idea is explicit in the formulation of the Turing test: a designer's challenge to build a machine that is indistinguishable from a human when interrogated (Turing, 1950). Since agency is a perceived attribute of humans, passing the Turing test implicitly also requires producing a perception of agency. Indeed, recent advancements in AI are often called breakthroughs because of the higher levels of perceived agency, including self-driving cars, language models that can write stories, or voice bots that can schedule appointments. Endowing an AI with perceived agency therefore reflects well not only on the ability of the AI but also that of its designer.

In Figure 2, these arguments are captured by the positive relationship from "Perceived agency of AI" to "Perceived ability of AI", and similarly by the positive relationship from "Perceived agency of AI" to "Perceived ability of AI's designer" (both indicated with a green, dashed arrow).

Mechanism 2: Shifting the Locus of Trust. The second addition to the baseline model in Figure 2 is to organize the relationships between trust in AI and (1) the ability, benevolence, and integrity perceptions of the AI, (2) those same perceptions of its designer, in the form of two different sets of beliefs, with agency attribution acting as a "tuning knob" that sets their relative strength. Trust in AI depends on trustworthiness perceptions about the AI and the designer of the AI, and we argue here that they are not always equally important. At one end, the relevant perceptions are mostly about the AI itself if significant levels of agency are attributed to it. At the other end, with low levels of agency attribution, the perceptions that matter are mostly about the AI's designer. While both sets of attributions may operate simultaneously, the extent to which trust in AI depends on the AI or on its designer, we argue, is tuned by the attribution of agency the human makes to the AI.

Because trust is a willingness to be vulnerable to another's actions, agency perceptions determine who will be seen as driving those actions. In philosophy, making judgements about a person's moral responsibility requires attributing agency to that person (Talbert, 2019). That is why children are held less responsible than adults in most legal systems (cf. Shariff et al., 2014). Similar reasoning holds for artificial agents. As AI is seen as more agentic, they will be seen as more responsible for their actions (Bigman et al., 2019). For example, people attribute more credit and blame to a robot that is more autonomous (Kim & Hinds, 2006; Van der Woerdt & Haselager, 2019). But responsibility for the actions of the AI does not solely rest with the AI itself. It also resides with its designer, in what has been referred to as "distributed agency" (Floridi, 2013; Taddeo, & Floridi, 2018). We argue that the relative importance placed on the AI versus its designer depends on the perceived agency of the AI.

Our argument here is consistent with a dual-process theory of cognition (Wason and Evans, 1974; Kahneman, 2003; Stanovich and West, 2000; summarized as System 1 versus System 2 thinking in Kahneman, 2011). If agentic perception is by default assumed to be high (since our trust-related interactions are for the most part with fellow humans), then it may be typical and automatic to trust based on the trustworthiness of the actor one is interacting with. However, with low levels of agentic cues, an explicit and deliberate consideration of the trustworthiness of the designer arises. Thus, the greater the perceptions of agency of the AI, the more important will be the trustworthiness of the AI relative to the trustworthiness of its designer in determining whether humans trust the AI.

This argument implies that perceived trustworthiness of the AI and of its designer both matter and that agentic attributions regulate their relative importance. For instance, consider a situation in which the AI is seen as highly trustworthy but its designer as not trustworthy. Then increasing perceived agency will increase trust in AI because the trustworthiness perceptions about the AI matter more than about the designer. Next, consider the opposite situation in which the designer is seen as high in trustworthiness but the AI as low in trustworthiness. In this case, making the AI appear more agentic will decrease trust in AI because of the increased importance of the trustworthiness perceptions about the AI compared to those about the designer.

Mechanism 3: Amplifying Betrayal Concerns. The third new idea in Figure 2 is that perceived agency of AI may affect the willingness to be vulnerable to an AI for any given level of its trustworthiness. For a human trustee, Bohnet and colleagues (Bohnet & Zeckhauser, 2004; Bohnet et al., 2008, 2010) have established that the anticipated psychological costs of trust betrayal hinders trust in the trustee, due to so-called betrayal aversion. This phenomenon is absent when playing a lottery against nature, which may have similar monetary costs of loss, but not the associated psychological costs of being betrayed.

Humans are less willing to be vulnerable when the source of risk is another human rather than nature, which distinguishes a trust decision from a risk decision. A main cause for this difference is the psychological cost of betrayal. Betrayal is defined as “a voluntary violation of mutually known pivotal expectations of the trustor by the trusted party (trustee)” (Elangovan & Shapiro, 1998: 548). Betrayal occurs when a trust decision goes wrong but not when a risk decision goes wrong. Betrayal costs are the psychological loss felt when one is betrayed by another party and are in addition to any material costs (Bohnet et al., 2008). Humans actively seek to avoid the possibility of being betrayed; they display betrayal aversion (Aimone & Houser, 2012). The existence of betrayal aversion has been widely documented (Bohnet & Zeckhauser, 2004; Butler & Miller, 2018; Lee et al., 2021), including in studies conducted in different countries across multiple continents (Bohnet et al., 2008, 2010). A neurological foundation for betrayal aversion has been identified (Lauharatanahirun et al., 2012, Aimone et al., 2014). Our close relatives, chimpanzees, also display betrayal aversion (Calcutt et al., 2019).

Betrayal aversion acts as a barrier to trust. Bohnet and colleagues (Bohnet & Zeckhauser, 2004, Bohnet et al., 2008, 2010) directly compare a trust game and a lottery. In the trust game, a trustor chooses whether to trust the human trustee or not. Without trust, the trustor obtains a certain payoff. With trust, the trustor obtains an uncertain payoff. It is higher than the certain payoff if the trustee honors the trust. It is lower if the trustee betrays the trustor (and it is in the trustee's interest to do so). In the lottery, the payoffs are the same: a certain, medium payoff and an uncertain, high or low payoff. If the trustor chooses the uncertain payoff, the actual payoff is determined by random chance (i.e., nature) and not by a human trustee. The consistent finding is that people are much more likely to choose the uncertain payoff in the lottery than in the trust game, for the same level of risk (see also Butler & Miller, 2018; Lee et al., 2021). This difference remains even after accounting for social preferences that could lead the trustor to avoid outcomes that are unequal compared to those of the trustee. In other words, betrayal aversion is a stronger form of concern for fairness. For these reasons, betrayal aversion makes humans less keen to be vulnerable to other humans than to nature.

Multiple lines of evidence tell us that perceived agency drives betrayal aversion (Butler and Miller, 2017; Lee et al., 2021). First, Bohnet and colleagues (Bohnet & Zeckhauser, 2004, Bohnet et al., 2008, 2010) compare a trust game also with a so-called risky dictator game. As in the trust game, a human player 1 chooses between a certain payoff (that is medium) or an uncertain payoff (that is either low or high). Also as in the trust game, a human player 2 benefits from "betrayal" if player 1 chooses the uncertain payoff. The only difference in this game is that player 2 lacks agency. Instead, the "betray" option is determined randomly. They find that humans are less willing to be vulnerable to agentic humans than to non-agentic humans. Second, Butler and Miller (2017) investigate the effect of human counter-party's intentionality using an adapted game. Specifically, the counter-party is forced to choose blindly between the honor and betray options, and the trustor is aware of this. When the counter-party's intentionality is taken away, betrayal aversion for the trustor is absent. Thus, perceived agency is a critical component of betrayal aversion. Put differently, it is not only that we get disutility from receiving an outcome that is perceived to be unfair relative to another's outcome (Rabin, 1993; Fehr & Schmidt, 1999), but additionally we get disutility because that unfair outcome is caused by an agentic actor. Thus, betrayal adds insult to injury, and the insult is a form of disutility that people will seek to avoid.

Whereas betrayal aversion is normally present with another human and absent with nature, its existence when interacting with an AI may be a matter of degree. Thus, we assume that betrayal aversion displayed toward agentic-seeming human actors can also manifest itself towards agentic-seeming non-human actors. The extent to which this assumption is valid is ultimately an empirical question. As we have argued above, AI is somewhere in between inanimate technology and a human in terms of perceptions of its agency. Enhancing the perceived agency of AI should therefore give rise to betrayal aversion concerns, because an agentic AI can actively betray you whereas a non-agentic AI cannot. The implication is that by itself higher agency perception can reduce trust in the AI because of the additional concern that betrayal would be painful, for any given level of trustworthiness behavior, relative to a non-agentic AI. An increase in how agentic an actor is perceived to be can simultaneously increase the importance of trustworthiness of that actor in determining whether to trust them, as well as lower the likelihood of trusting them because the expected cost of betrayal increases.

This consideration of increasing levels of agentic attribution is not normally necessary when dealing with either humans (because agency is presumed to exist) or a non-agentic technology (because it is presumed absent in the technology but present in its human designer). But with AI, particularly the kind that seems to act autonomously and opaquely through deep learning, there are degrees to which attributions of agency are made. This, we propose, plays a large role in whether humans trust AI. Thus, holding constant the perceived trustworthiness of the AI and its designer, the greater the perception of agency of the AI, the lower the trust in AI will be.

Implications of the Model

This model of perceived agency of AI and trust in it has several implications that can be tested in future research. We focus on three sets. First, we focus on static trust implications. Agency perceptions affect a user's initial trust in AI. Second, we focus on dynamic trust implications. Agency perceptions affect how a user's trust develops through interacting with the AI. Third, we focus on the implications of dynamic agency perceptions. Through interacting with an AI, a user may change not only its level of trust in AI but also its agency perceptions about the AI.

Static trust implications. The model specifies three mechanisms through which agency perceptions affect a user's initial trust in AI. In isolation, each mechanism has the following implications. Greater perceived agency enhances trust in AI through strengthening ability perceptions (mechanism 1), enhances trust through shifting the locus of trust if the AI is seen as trustworthy but not its designer and reduces it in the opposite situation (mechanism 2), or reduces trust through amplifying betrayal concerns (mechanism 3). Together, these three mechanisms can thus imply that greater agency perceptions about an AI lead to lower or higher human trust in AI. The model thus indicates that increasing agency perceptions are likely to *decrease* trust in the AI, when betrayal concerns are vivid or the AI is perceived as untrustworthy (i.e., has lower ability, benevolence and integrity) compared to its designer. In contrast, increasing agency perceptions *increase* trust when the AI is perceived as more trustworthy compared to its designer.

Dynamic trust implications. Agency perceptions affect not only a user's initial trust in AI but also how that trust develops over multiple interactions. We explore two arguments about changes in trust. The first is that agency perceptions influence the *direction* of change. The second is that agency perceptions influence the *speed* of change.

Perceived agency affects the direction of trust development because of its role in initial ability perceptions. Repeated interactions with a counterparty yield information about the other's trustworthiness (Anderson & Weitz, 1989; Shapiro et al., 1992; Lewicki & Bunker, 1995). If initial perceptions about the other's trustworthiness are too low, then over time the trustee's actions and outcomes will lead a trustor to revise its trustworthiness perceptions upward, leading to greater trust. In contrast, if initial expectations are too high, then trustworthiness perception and hence trust tend to decline over time (Vanneste et al., 2014). If initial expectations are in line with actual trustworthiness, then on average trust remains the same. Thus, initial trustworthiness perceptions influence the direction of trust development.[7]

[7] This literature (e.g., Anderson & Weitz, 1989; Shapiro et al., 1992; Lewicki & Bunker, 1995; Vanneste et al., 2014) assumes that over time trustworthiness perceptions are updated towards the level of actual trustworthiness. If a user exhibits confirmation bias, then that updating is slowed.

Perceived agency enhances initial ability perceptions (as per mechanism 1 in the model). and repeated interactions may lead to fewer trust increases (because actual ability is less likely to exceed perceived ability) or more trust decreases (if initial ability perceptions were too optimistic). For instance, a user who perceived an AI as not very agentic (and thus not very able) would increase its trust after a series of good outcomes by the AI. In contrast, a different user who perceived the same AI as highly agentic (and therefore highly able) may not increase its trust in the AI after a similar series of good outcomes. It is interesting to note that AI often suffers from unrealistically positive expectations (Burton et al., 2020), which enhanced agency perceptions would only amplify. In contrast, if initial trustworthiness expectations are pessimistic despite agency perceptions, then trust in AI will increase over time (but less so than without agency perceptions). The implication is that regardless of whether initial ability perceptions are too high or too low, holding constant the actual trustworthiness of the AI, perceived agency dampens the trajectory of trust development: it makes increases less likely and decreases more likely.

Perceived agency affects not only the direction of trust development but also its speed. In general, trustee behavior that signals trustworthiness should enhance trust. Conversely, behavior that signals less trustworthiness should do the opposite. However, the updating rate is itself contingent on the agency attributions that the trustor makes. Trust is harder to build despite trustworthy actions if they can be attributed to constraints rather than agency (Puranam & Vanneste, 2009). In the course of any exchange relationship, opportunities typically arise for the parties to engage in either trustworthy or untrustworthy behavior. When agentic attributions are made of the trustee, trustworthy behavior is likelier to be recognized as such, leading to an updated and increased level of trust (Anderson & Weitz, 1989; Parkhe, 1993; Ring & Van de

Ven, 1992). However, the absence of such an attribution can impede trust formation, since trustworthy behavior is less likely to be attributed to the actor than to the situational constraints (Malhotra & Murnighan, 2002; Strickland, 1958). The necessity of agency attributions for assigning and updating beliefs about trustworthiness is recognized by philosophers of ethics, who typically hold that one cannot attribute moral responsibility without agency (e.g., Gray et al., 2007; Bigman et al., 2019). Therefore, weaker agency attributions should lead to a slower rate of trust development for the same level of trustworthy behavior by the trustee. Conversely, agency attributions to the trustee should also make the trustor more sensitive to untrustworthy behavior by the former, leading to a more rapid erosion of trust.

Finally, with increasing agency perceptions of an AI, the locus of trust shifts from the designer towards the AI (mechanism 2). In this case, the AI's perceived trustworthiness matters more and that of the designer less. Consequently, if the AI is seen as more agentic, then outcomes lead to faster updates in trustworthiness perceptions about the AI and slower updates to those about its designer. Hence, we may expect that an AI algorithm's failure is more damaging for the reputation of the company that made it if the algorithm is seen as less agentic.

Dynamic agency perceptions implications. A user may change not only its level of trust in an AI but also its agency perceptions about the AI over multiple interactions. In a first interaction, a user may assess how agentic an AI is. If it is difficult to directly assess the degree of agency, then a user may base initial agency perceptions on prior interactions with other AI. In other words, just as people have trustworthiness perceptions about others they have never met (Nannestad, 2008), we know that they may also have agency perceptions about AI they have never used (Gray et al., 2007).

Over further interactions, however, perceptions of how agentic the AI is may change. Over multiple interactions with a modern, connectionist AI, one may come to observe it changing its behavior over time, for instance, by adapting its behavior to previous interactions. This ability to change behavior without the apparent direct intervention of a human designer can lead to greater perceptions of autonomy, leading to an overall increase in perceptions of agency. Multiple such interactions offer an opportunity to revise upward the beliefs about the system's autonomy.

However, multiple interactions may also lead to a greater sense of understanding of how and why the AI acts as it does. Even without a detailed understanding of the underlying machinery, we may come to form "theories of mind" (Gray et al, 2007; Gray and Wegner, 2012) about how and why the AI is acting in particular ways. These theories may be illusory; but to the extent that they are a product of repeated interaction and a greater sense of familiarity, they could lower perceptions of opacity and therefore decrease perceptions of agency about the AI (Bigman et al., 2019). Multiple interactions can therefore have opposing effects on the perceptions of agency about an AI system held by its human interactor. It may increase because of enhanced perception of autonomy, or it may decrease because of a diminished sense of opacity. The relative magnitudes of these effects will likely vary across contexts.

How do agency perceptions changing over time affect trust in AI? The model offers several suggestions for how to investigate this. Most fundamentally, it indicates that the strength of the three mechanisms may vary over repeated interactions with an AI. With agency perceptions increasing over time, betrayal concerns become more important, the designer moves towards the background, and ability perceptions are heightened. In contrast, with agency perceptions decreasing over time, betrayal concerns become less important, the designer becomes more important, and ability perceptions are depressed. However, it is also worth noting that the opportunity for further interactions (and subsequent changes in agency perceptions) itself depends on satisfactory past interactions (i.e., trust being justified), which in turn depends on the levels of trust that arose from past perceptions of agency. This dynamic system can have both positive and negative feedback loops and would require a formal analysis under specific assumptions for specific predictions to be derived. Taken together, we can say that varying agency perceptions means not only that people trust different AI differently, but also that the level of trust in a given AI may change as a user learns more about the AI and its agentic nature.

Contingent Predictions

To facilitate testing the implications of the model, we provide several directional predictions for exogenous contingencies (see Table 1). We consider the following contingencies. Task complexity is the difficulty of successfully completing a task because it comprises many interdependent components. Designer reputation refers to how the user perceives the AI's creator or company (Rindova et al., 2005). Potential loss addresses the possible negative consequences, either financially or personally, when things go wrong while relying on the AI. Application novelty is about how new or unprecedented the use case of the AI is. A trust breach occurs when the AI takes an action that violates the user's trust in it (Kim et al., 2006). Designer scandal refers to a controversy or wrongdoing involving the AI's creator or the associated company. Explainability pertains to the AI's capacity to articulate or justify its actions and decisions. Lastly, updating describes the AI's ability to refine its algorithms or models based on new data.

[[INSERT TABLE 1 ABOUT HERE]]

INTERVENTIONS THAT CHANGE TRUST IN AI

Practitioners and academics are interested in trust in AI because a lack of trust may prevent the adoption or use of a beneficial AI system. Conversely, unjustified trust may lead to harmful adoption. At first glance, AI is no different than other technologies, for which adoption critically hinges on confidence (McKnight et al., 2011). Research on technology adoption has sought to understand the interventions that influence confidence (Wu et al., 2011) and so has the research specifically focused on AI technologies and trust (Glikson & Woolley, 2020). We argue however that a key difference is that interventions meant to stimulate trust in AI may also affect agency perceptions of the AI, which in turn may influence trust levels in line with, or opposite to, the intended effect.

To highlight these implications of agency perceptions, we discuss several interventions frequently used to affect trust, including enhancing an AI's autonomy (i.e., giving it a greater ability to decide without consulting a human), improving its reliability (i.e., increasing its performance), providing greater transparency (i.e., explaining the logic for its decisions and actions), and anthropomorphizing it (i.e., making it more human-like in appearance and capabilities). Although these interventions are not new, we use our framework to provide an integrated understanding of them, focusing on the implications for agency perceptions.

Based on our theory, we can predict the conditions under which increasing agency of an AI will also lead to an increase in human trust in it. First, our model predicts that this will depend on whether the AI's attained perceived trustworthiness exceeds that of the designer. Conversely, it should be possible to increase human trust in the AI by downplaying its agency, and instead enhancing perceptions of the designer's trustworthiness. The success of this approach depends on whether the perceptions of the designer's trustworthiness are already high or can be improved. A survey among the US public indicates that some AI companies certainly have scope for improving how trustworthy they are perceived to be (Zhang et al., 2021). Second, if the agentic perception of the AI is exogenous and cannot be changed easily, then our model also predicts when it will be more effective to invest in enhancing the perceived trustworthiness of the AI (i.e., when its perceived agency is high) vs. its designer (when the AI's perceived agency is low). Third, the framework incorporates perceptions not only about agency but also about the AI's and designer's trustworthiness. It makes clear that the effectiveness of interventions to enhance trustworthiness perceptions when directed at the designer vs. the AI depends on the level of perceived agency. We consider several interventions that have been examined in prior literature through the lens of our framework (see Table 2).

[[INSERT TABLE 2 ABOUT HERE]]

The first intervention we consider is *autonomy* or giving the AI a greater ability to decide on a course of action without consulting a human (Kim & Hinds, 2006; Waytz et al., 2014). For example, consider the difference between an AI making investment recommendations on asset allocations and one that changes asset allocations depending on market conditions without human intervention. In the latter case, agency perception of the AI should be higher (Kim & Hinds, 2006). We do not anticipate an effect on benevolence or integrity perceptions, as increasing autonomy does not necessarily mean that an AI is more or less likely to do what is good for the human. On the other hand, ability perceptions may increase because of enhanced autonomy, as we have argued under our first mechanism. However, since this also places more emphasis on the perceived trustworthiness of the AI as well as raises concerns about betrayal aversion, an increase in autonomy for the AI is by no means guaranteed to increase trust in it by humans. It is likelier to do so if the trustworthiness of the AI's designer is lower than that of the

AI itself (which is made unlikely by the fact that increased autonomy of the AI will also raise the perceived ability and therefore trustworthiness of the designer), or if the increase in AI's trustworthiness (via an increase in its perceived ability) is large enough to offset the increased concerns about betrayal aversion.

In contrast, the second intervention we consider—increasing *reliability* or increasing the performance of the AI system—should have unambiguously positive effects on trust. This intervention should enhance ability perceptions of the AI (as well as its designer) but does not necessarily affect benevolence or agency perceptions about it. Therefore, trust in the AI is expected to go up, which is consistent with the current empirical evidence (Glikson & Woolley, 2020). Increasing the reliability of the AI should therefore be a more robust intervention than increasing its autonomy if the goal is to enhance human trust in AI.

The third intervention we consider is *transparency* or explaining the logic of an AI's decisions or actions. Many AI algorithms based on connectionism are black boxes in that a human does not see the reasons for a decision. Furthermore, they often rely on non-linear function approximation, which makes explaining their working in an intuitive manner to humans difficult (Lipton, 2018). This may make users hesitate to adopt the algorithm's recommendations, or in some cases to over-rely on them (Bussone et al, 2015; for a review of explainable AI applications in medicine and the challenge of generating trust in them, see Tjoa & Guan, 2020). Efforts are increasingly made to provide explanations of AI's decision making or so-called explainable AI (XAI). Benevolence or integrity perceptions are unlikely to be systematically affected by explainability (the explanation of a decision may make the decision seem more or less benevolent or honorable to the human). On the other hand, ability perceptions are unambiguously enhanced if decisions are seen as less random. Therefore, the first effect of explainability should be to increase the perceived trustworthiness of AI via enhanced ability perceptions.

A second effect of transparency is to diminish perceptions of agency, because an increase in transparency implies a reduction in opacity. Making AI more explainable may thus detract from its capacity to be treated as an agent. Therefore, trust in AI would go up with increased transparency only if the designer is deemed more trustworthy than the AI, else it might go down. Combining both effects, trust in AI could increase or decrease with transparency because betrayal aversion concerns are lessened, and ability perceptions for both the AI and its designer are enhanced, but the importance of these perceptions about the AI themselves are reduced.

The last intervention we consider is *anthropomorphizing* or making AI more human-like, in terms of both physical appearance and human mind like qualities (Epley et al., 2007). Examples include giving a name, gender, and voice to autonomous vehicles (Wyatz et al., 2014). A consequence of anthropomorphizing is higher agency perceptions of the AI because humans are usually credited with agency. This should activate the betrayal aversion mechanism. To the extent that the AI comes to resemble a human, benevolence perceptions may also increase (Glikson & Woolley, 2020). One mechanism behind this is homophily, whereby others like us are trusted more (Foddy & Yamagishi, 2009), or to put it negatively, speciesism, whereby other species (or types of being) are trusted less. The effect on ability perceptions is less clear. If anthropomorphizing leads to ability perception of an AI similar to that of a human, then ability perception of AI may increase or decrease depending on whether a human was seen as more or less able. Taken together, anthropomorphizing need not increase trust in AI, even though that is often a key goal of anthropomorphizing. Whereas benevolence perceptions may be increased, this could be offset by an increase in betrayal aversion concerns.

In sum, interventions that increase perceptions of AI agency—such as letting a robot decide on its own course of action (Kim & Hinds 2006) (autonomy), providing explanations of how the AI works (transparency) or giving a name, gender, and human voice to an autonomous vehicle (Waytz et al. 2014) (anthropomorphization)—may increase or lower trust in AI depending on the contingencies we have set out above.

DISCUSSION

Existing theories of interpersonal trust often assume that the trustee has a significant degree of agency since trust is undefined unless the trustee has the opportunity to act freely. This assumption is challenged by the phenomenon of human trust in AI. Perceptions of agency about AI need not be discrete but can take on a range of values, substantially changing how trust unfolds. Our arguments imply that designing AI systems to produce greater agency perceptions is neither necessary nor sufficient for producing greater human trust in such systems. Specifically, if an AI system is perceived to have no agency at all, our theory predicts that the locus of trust will shift to the designer of the system. In contrast, if it is perceived as fully agentic, the designer's attributes fade into the background: the more lifelike the puppet, the less likely we are to notice the strings (and the puppeteer). At intermediate levels of perceived agency, the system's and the designer's attributes both influence the decision to trust.

Depending on beliefs about the relative trustworthiness of the designer and the AI, this implies that increasing agency perceptions may increase or decrease the likelihood that humans will trust an AI system. We use this reasoning to offer an explanation for why interventions designed to increase the human-like appearance of AI (which should in general increase the chances of the system meeting the Turing test) may in fact produce lower chances of human trust in the AI. The concept of the “uncanny valley” has become popular in describing the non-monotonic relationship between the human likeness of a designed object (a robot, a virtual avatar, or even a doll) and the aesthetic response of humans to it. Humans seem to display an increasing degree of liking as the object appears more human up to a point, beyond which there is a steep decline in liking for the quite human-like object, followed again by an increase in liking as the object becomes near-indistinguishable from humans (Mori, 1970). Our analysis points to the likelihood of non-monotonic relationships between agency perceptions of AI and human trust in AI, which need to be empirically investigated. While we cannot say if the relationship will take the same shape as the “uncanny valley”, we offer a theory about the underlying forces that can help guide the search for the shape of this relationship.

Open Questions and Opportunities for Further Research

We have been agnostic in this paper as to whether (or when) one should engineer greater human trust in AI systems. This normative question should precede the question of *how* to engineer such trust, but it is outside the scope of our analysis. Instead, we have focused on the consequences of how agentic an AI is perceived to be for human trust in that AI. We believe the answer as to whether human trust in AI should be enhanced will vary significantly by context. For instance, when there is good reason to believe that trust deficits are preventing dependence on the AI that will produce benefits for human users, our analysis suggests interventions that can overcome this trust deficit by enhancing or suppressing the perceived agency of the AI. In contrast, where there is risk of over-reliance on AI as has been documented in some medical applications (Tjoa & Guan, 2015; Sujan et al., 2019), and has become salient with the extremely rapid diffusion of large language models like ChatGPT, designers may consciously alter agency perceptions in order to lower the trust that humans are likely to place in the AI system.

More generally, many today are concerned that short-term willingness to trust and depend on AI may become long-term over-reliance and human helplessness. Such concerns often accompany the self-reinforcing dynamic of specialization and have been previously articulated in the context of firms outsourcing to other firms, or of one country's economy becoming dependent on another's through offshoring. When such concerns are deemed legitimate, policy makers may legislate against design features that affect agency perceptions about AI in a manner that produces high levels of willingness to trust by humans. Such decisions require consideration of factors that are outside the scope of the current paper, but they suggest a rich avenue for further development, including on the timing of any such regulation.

Our treatment of agency perceptions in trust has taken a ceteris paribus approach to other variables that could be relaxed in future work. For instance, it has long been recognized that uncertainty is crucial for trust (Gambetta, 1988; Yamagishi & Yamagishi, 1994). We have focused on the perceived agency of the trustee as a crucial source of that uncertainty. As Hardin (2002: 12) noted:

> "*...one might say trust is embedded in the capacity or even need for choice. Giving people overwhelmingly strong incentives seems to move them toward being deterministic actors with respect to the matters at issue. (That is one reason romantics detest rationality.) At the other extreme, leaving them with no imputable reasons for action makes it impossible to trust them in many contexts.*"

To be sure, uncertain outcomes can result even when relying on a non-agentic system through uncontrollable, unforeseen, and poorly understood causes. However, the trustor's interpretation of the uncertainty in the situation will vary depending on the perceived agency of the trustee. Therefore, our focus has been on the implications of varying perceptions of agency for a given level of uncertainty in outcome. Further development could investigate the effects of agency perceptions on uncertainty in addition to the mechanisms articulated in this paper.

Our analysis has also treated the benevolence of the AI and its designer as exogenous factors not directly affected by perceptions of agency about the AI. Just because an entity is seen as agentic does not imply that it will be expected to act in a more-or-less benign manner. The popular meme of "killer AI" associates increased agency of AI with a decline in benign intentions towards humans, but this is only true if there is misalignment of goals. In our social life, we routinely encounter friends and family members who are both highly agentic and benign. Therefore, there is no a priori reason to assume that changes in agency perception should affect perceptions of benevolence or integrity in one direction or the other. Our results cannot be seen as indicating the impact of changes in agency perception on human willingness to be vulnerable to AI when agency perceptions have no direct impact on perceived benevolence of AI. In a particular context, if we expect an effect of agency perception on the perceived benevolence or integrity of AI, our framework can easily be extended to assess the impact of changed agency perceptions on human trust in AI in that context.

Another area in which our framework could be extended is to explore if the different components of perceived trustworthiness (ability, benevolence, and integrity) are amplified differently with higher agency perceptions (mechanism 2). For what actions or outcomes do agency perceptions matter most for attributions about ability, benevolence, or integrity? For example, consider a self-driving car that skillfully avoids an accident. If the self-driving car is perceived as agentic, then trustworthiness perceptions about the car might be enhanced more than about its designer. If the car narrowly avoided a rock that fell on the road, then ability perceptions might be most affected. If the car manages to narrowly avoid a pedestrian by swerving off the road and in the process hits a rock, then integrity perceptions might be most affected (updated upwards). Thus, specific actions or outcomes may influence which perceptions are amplified most.

Our model analyzes the role of perceived agency in trust in AI, but it does not explore the implications for different types of AI. We believe that understanding the role of different types of AI is an exciting area for future research. For example, AI can manifest itself as a robot, a virtual agent, or embedded in another device (e.g., in a computer) (Glikson & Woolley, 2020). Its manifestation will not only affect the creation of agency perceptions (Epley et al., 2007; Bartneck et al., 2009), but may also influence how those perceptions affect trust.

In our analysis, we focus on situations where (1) the designer is human or, more typically, a collection of humans (i.e., an organization, see Zaheer et al., 1998; Vanneste, 2016); (2) the designer's identity is known sufficiently to make attributions about; and (3) the designer can be treated as a single entity. One could extend this analysis in three ways. First, a designer could be artificial as when an AI algorithm proposes a molecular structure for a new medicine (a non-AI technology) or when it designs another AI algorithm (an AI technology) (e.g., AutoML or automated machine learning). Second, a user may not always know the designer's identity. Not knowing the designer may make users reluctant to use the technology (Kuester et al., 2018; Konya-Baumbach et al., 2019). Relatedly, calls have been made for the designer to be identified to a user to enhance trust in an AI (Andras et al., 2018). Third, one could think of the designer not as a single entity but as a chain of (artificial or human) designers. An opportunity exists for future research to investigate each of these possibilities. For instance, on the last point, we might speculate that designer's attributes may play a stronger role in shaping trust in the system if they are themselves seen as more agentic.

Finally, while our focus has been on the case of human trust in AI, our arguments about the role of variable degrees of agency perception on trust also apply to interpersonal trust. For instance, once we allow for variable degrees of agency, it is possible to ask about the differences between how a customer may trust an employee (perceived as having limited agency) vs. the owner of a business. If the goal is to enhance customer trust in the employee, our argument implies that the perceived autonomy of the employee in terms of delegation of decision rights in a public and credible manner (Stea et al., 2015; Vergne, 2020) would become an important design variable to be tuned based on the relative trustworthiness perceptions of the employee and the owner of the business.

A related idea is that extensive governance mechanisms may adversely influence trust (Poppo & Zenger, 2002; Mayer & Argyres, 2004). For example, trust development may be limited when the other party's benign actions are attributed not to the good intentions of the other party but to the governance mechanisms that force agents to behave well (Strickland, 1958; Malhotra & Murnighan, 2002; Puranam & Vanneste, 2009). In exchange relationships that occur in the context of binding contracts (Malhotra & Murninghan, 2002), the development of trust over time may be impeded if there is limited scope for the partners to potentially engage in opportunistic behavior (though not zero scope since all contracts are incomplete). This notion has been referred to as the indirect crowding-out effect of contracts on trust, distinct from the direct effect of imposing a contract that may signal distrust (Puranam & Vanneste, 2009). Therefore, the extensiveness of contractual clauses becomes a design variable that can affect ex-ante trust as well as trust development between contracting parties. Working out the theoretical and empirical implications of restrictions on human agency seems a fruitful avenue for expanding our understanding of trust.

CONCLUSION

Perceptions of agency play an implicit role in theories of interpersonal trust, since we usually assume the humans we interact with have a considerable degree of agency. Indeed, this is what distinguishes interpersonal trust from confidence in an inanimate technology (which in turn may be founded on trust in its human designer). Modern AI systems are different in that they may be perceived to have varying degrees of agency that lie between that of inanimate technology and an agentic human, however. This should affect human willingness to be vulnerable to AI systems, and we have argued that three distinct mechanisms are relevant: agency perceptions lead to a change in ability perceptions, a shift in the locus of trust, and betrayal aversion. The joint operation of these mechanisms suggests that increasing perceptions of how agentic an AI system is, need not always produce a greater human willingness to trust: success at meeting the Turing test may go hand in hand with decreased human trust of AI.

REFERENCES

Agrawal A., Gans J., & Goldfarb A. 2018. ***Prediction Machines: The Simple Economics of Artificial Intelligence***. Boston, MA: Harvard Business Review Press.

Aimone, J. A., & Houser, D. 2012. What you don't know won't hurt you: A laboratory analysis of betrayal aversion. ***Experimental Economics***, 15: 571-588.

Aimone, J. A., Houser, D., & Weber, B. 2014. Neural signatures of betrayal aversion: An fMRI study of trust. ***Proceedings of the Royal Society B: Biological Sciences***, 281: 20132127.

Anderson, E., & Weitz, B. 1989. Determinants of continuity in conventional industrial channel dyads. ***Marketing Science***, 8: 310–323.

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., Urquhart, N., & Wells, S. 2018. Trusting intelligent machines: Deepening trust within socio-technical systems. ***IEEE Technology and Society Magazine***, *3*7: 76-83.

Banks, J. 2019. A perceived moral agency scale: development and validation of a metric for humans and social machines. ***Computers in Human Behavior***, 90: 363-371.

Barocas, S., & Selbst, A. D. 2016. Big data's disparate impact. ***California Law Review***, 104: 671-732.

Baumeister, R. F., & Monroe, A. E. 2014. Recent research on free will: Conceptualizations, beliefs, and processes. ***Advances in Experimental Social Psychology***, 50: 1-52.

Belanche, D., Casaló, L. V., Schepers, J., & Flavián, C. 2021. Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The Humanness‐Value‐Loyalty model. ***Psychology & Marketing***, 38: 2357-2376.

Benbasat, I., & Wang, W. 2005. Trust in and adoption of online recommendation agents. ***Journal of the Association for Information Systems***, 6: 72-101.

Bengio, Y., Lecun, Y., & Hinton, G. 2021. Deep learning for AI. ***Communications of the ACM***, 64: 58-65.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. 2019. Holding robots responsible: The elements of machine morality. ***Trends in Cognitive Sciences***, 23: 365-368.

Bhattacharya, R., Devinney, T. M., & Pillutla, M. M. 1998. A formal model of trust based on outcomes. ***Academy of Management Review***, 23: 459-472.

Boden, M. A. (1996). Autonomy and artificiality. In M. A. Boden (Ed.), T***he Philosophy of Artificial Life***, 95-108. Oxford, U.K.: Oxford University Press.

Bohnet, I., & Zeckhauser, R. 2004. Trust, risk and betrayal. ***Journal of Economic Behavior & Organization***, 55: 467-484.

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. 2008. Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. ***American Economic Review***, 98: 294-310.

Bohnet, I., Herrmann, B., & Zeckhauser, R. 2010. Trust and the reference points for trustworthiness in Gulf and Western countries. ***Quarterly Journal of Economics***, 125: 811-828.

Brave, S., Nass, C., & Hutchinson, K. 2005. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. ***International Journal of Human-Computer Studies***, 62: 161-178.

Burton, J. W., Stein, M. K., & Jensen, T. B. 2020. A systematic review of algorithm aversion in augmented decision making. ***Journal of Behavioral Decision Making***, 33: 220-239.

Bussone, A., Stumpf, S., & O'Sullivan, D. 2015. The role of explanations on trust and reliance in clinical decision support systems. ***2015 International Conference on Healthcare Informatics***, 160-169.

Butler, J. V., & Miller, J. B. 2018. Social risk and the dimensionality of intentions. ***Management Science***, 64: 2787-2796.

Calcutt, S. E., Proctor, D., Berman, S. M., & De Waal, F. B. 2019. Chimpanzees (Pan troglodytes) are more averse to social than nonsocial risk. ***Psychological Science***, 30:(1), 105-115.

Chadderdon, G. L. 2008. Assessing machine volition: An ordinal scale for rating artificial and natural systems. ***Adaptive Behavior***, 16: 246-263.

Clarke, R. 1995. Indeterminism and control. ***American Philosophical Quarterly***, 32: 125-138.

Colquitt, J. A., Scott, B. A., & LePine, J. A. 2007. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. ***Journal of Applied Psychology***, 92: 909-927.

Connelly, B. L., Crook, T. R., Combs, J. G., Ketchen Jr, D. J., & Aguinis, H. 2018. Competence-and integrity-based trust in interorganizational relationships: Which matters more? ***Journal of Managemen***t, 44: 919-945.

Csaszar, F., & Steinberger, T. 2021. Organizations as artificial intelligences: The use of artificial intelligence analogies in organization theory. ***Academy of Management Annals***. 16: 1-37.

Culley, K. E., & Madhavan, P. 2013. Trust in automation and automation designers: Implications for HCI and HMI, ***Computers in Human Behavior***, 29: 2208-2210.

Davis, F. D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. ***MIS Quarterly***, 13: 319-340.

Dennett, D. C. 1991. ***Consciousness Explained***. Boston, MA: Little Brown.

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. ***Journal of Experimental Psychology: Applied***, 22: 331-349.

Dietvorst, B. J., Simmons, J. P., & Massey, C. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. ***Journal of Experimental Psychology: General***, 144: 114-126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. ***Management Science***, 64: 1155-1170.

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. 2019. Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. ***Computers in Human Behavior***, 90: 357-362.

Elangovan, A. R., & Shapiro, D. L. 1998. Betrayal of trust in organizations. ***Academy of Management Review***, 23: 547-566.

Epley, N., Waytz, A., & Cacioppo, J. T. 2007. On seeing human: A three-factor theory of anthropomorphism. ***Psychological Review***, 114: 864-886.

Franklin, S., & Graesser, A. 1997. Is it an Agent, or just a Program? A Taxonomy for Autonomous Agents. In ***Intelligent Agents III, Agent Theories, Architectures, and Languages (ATAL 1996)***, 21-35.

Fehr, E., & Schmidt, K. M. 1999. A theory of fairness, competition, and cooperation. ***Quarterly Journal of Economics***, 114: 817-868.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. 2019. A taxonomy of social cues for conversational agents. ***International Journal of Human-Computer Studies***, 132: 138-161.

Fiore, S. M., Wiltshire, T. J., Lobato, E. J., Jentsch, F. G., Huang, W. H., & Axelrod, B. 2013. Toward understanding social cues and signals in human–robot interaction: Effects of robot gaze and proxemic behavior. ***Frontiers in Psychology***, 4: 859.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. ***Journal of Personality and Social Psychology***, 82: 878–902.

Fiske, S. T., Cuddy, A. J., & Glick, P. 2007. Universal dimensions of social cognition: Warmth and competence. ***Trends in Cognitive Sciences***, 11: 77-83.

Floridi, L. 2013. Distributed morality in an information society. ***Science and Engineering Ethics***, 19: 727-743.

Foddy, M., & Yamagishi, T. 2009. Group-based trust. In K. S. Cook, M. Levi, & R. Hardin (Eds.), ***Whom Can We Trust? How Groups, Networks, and Institutions Make Trust Possible*** (pp. 17-41). New York, NY: Russell Sage.

Fogg, B. J., & Nass, C. 1997. How users reciprocate to computers: an experiment that demonstrates behavior change. ***Human Factors in Computing Systems***, 331-332.

Gambetta, D. 1988. Can we trust trust? In Gambetta, D. (Ed.), ***Trust: Making and Breaking Cooperative Relations***, 213-237. Cambridge, MA: Basil Blackwell.

Glass, A., McGuinness, D. L., & Wolverton, M. 2008. Toward establishing trust in adaptive agents. In ***Proceedings of the 13th International Conference on Intelligent User Interfaces***, 227-236.

Glikson, E., & Woolley, A. W. 2020. Human trust in artificial intelligence: Review of empirical research. ***Academy of Management Annals***, 14: 627-660.

Gottfredson, L. S. 1997. Why g matters: The complexity of everyday life. ***Intelligence***, 24: 79-132.

Gray, H. M., Gray, K., & Wegner, D. M. 2007. Dimensions of mind perception. ***Science***, 315: 619-619.

Gulati, S., Sousa, S., & Lamas, D. 2018. Modelling trust in human-like technologies. In ***Proceedings of the 9th Indian conference on Human Computer Interaction*****,** December: 1-10.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. 2011. A meta-analysis of factors affecting trust in human-robot interaction. ***Human Factors***, 53: 517-527.

Hardin, R. 2002. ***Trust and Trustworthiness***. New York, NY: Russell Sage Foundation.

Haugeland, J. 1985. ***Artificial Intelligence: The Very Idea***. Cambridge, MA: MIT Press.

Hengstler, M., Enkel, E., & Duelli, S. 2016. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. ***Technological Forecasting and Social Change***, 105: 105-120.

Higgins, M. C., & Gulati, R. 2003. Getting off to a good start: The effects of upper echelon affiliations on underwriter prestige. ***Organization Science***, 14: 244-263.

Höddinghaus, M., Sondern, D., & Hertel, G. 2021. The automation of leadership functions: Would people trust decision algorithms? ***Computers in Human Behavior****,* 116: 106635.

Hoff, K. A., & Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. ***Human Factors***, 57: 407-434.

Hornik, K., Stinchcombe, M., & White, H. 1989. Multilayer feedforward networks are universal approximators. ***Neural Networks*****,** 2: 359-366.

Hu, Q., Lu, Y., Pan, Z., Gong, Y., & Yang, Z. 2021. Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants. ***International Journal of Information Management***, 56: 102250.

Kahneman, D. 2003. A perspective on judgment and choice: mapping bounded rationality. ***American Psychologist***, 58: 697.

Kahneman, D. 2011. ***Thinking, Fast and Slow***. New York: Girard & Giroux.

Kallus, N., Mao, X., & Zhou, A. 2022. Assessing algorithmic fairness with unobserved protected class using data combination. ***Management Science***, 68: 1959-1981.

Kane, R. 1996. ***The Significance of Free Will***. New York, NY: Oxford University Press.

Kim, A., Cho, M., Ahn, J., & Sung, Y. 2019. Effects of gender and relationship type on the response to artificial intelligence. ***Cyberpsychology, Behavior, and Social Networking***, 22: 249-253.

Kim, D. J., Ferrin, D. L., & Rao, H. R. 2008. A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. ***Decision Support Systems***, 44: 544-564.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. 2006. When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. ***Organizational Behavior and Human Decision Processes***, 99: 49-65.

Kim, T., & Hinds, P. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. ***The 15th IEEE International Symposium on Robot and Human Interactive Communication***, 80-85.

Konya-Baumbach, E., Schuhmacher, M. C., Kuester, S., & Kuharev, V. 2019. Making a first impression as a start-up: Strategies to overcome low initial trust perceptions in digital innovation adoption. ***International Journal of Research in Marketing***, 36: 385-399.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. ***PLOS One***, 3: e2597.

Kuester, S., Konya-Baumbach, E., & Schuhmacher, M. C. 2018. Get the show on the road: Go-to-market strategies for e-innovations of start-ups. ***Journal of Business Research***, 83: 65-81.

Lankton, N. K., McKnight, D. H., & Tripp, J. 2015. Technology, humanness, and trust: Rethinking trust in technology. ***Journal of the Association for Information Systems***, 16: 880-918.

Lauharatanahirun, N., Christopoulos, G. I., & King-Casas, B. 2012. Neural computations underlying social risk sensitivity. ***Frontiers in Human Neuroscience***, 6, 213.

Laumer, S., Maier, C., & Gubler, F. 2019. Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. In ***Proceedings of the 27th European Conference on Information Systems (ECIS)***.

Lee, J., & Moray, N. 1992. Trust, control strategies and allocation of function in human-machine systems. ***Ergonomics***, 35: 1243-1270.

Lee, J. E. R., & Nass, C. I. 2010. Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In D. Latusek & A. Gerbasi (Eds.), ***Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives***, IGI Global: Hershey, PA.

Lee, J. D., & See, K. A. 2004. Trust in automation: Designing for appropriate reliance. ***Human Factors***, 46: 50-80.

Lee, M., Lucas, G., Mell, J., Johnson, E., & Gratch, J. 2019. What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In ***Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents***: 38-45.

Lee, S. Y., Kinias Z., & Vanneste, B. S. 2021. In Groups We Trust: Lower Betrayal Aversion Toward a Group than Toward an Individual. ***Working Paper***, available at https://ssrn.com/abstract=3983438

Legg, S., & Hutter, M. 2007. Universal intelligence: A definition of machine intelligence. ***Minds and Machines***, 17: 391-444.

Lewicki, R. J., & Bunker, B. B. 1995. Trust in relationships: A model of development and decline. In B. B. Bunker, & J. Z. Rubin (Eds.), ***Conflict, Cooperation and Justice***: 133–173. San Francisco, CA: Jossey-Bass.

Lindebaum, D., Vesa M., & Den Hond, F. 2020. Insights from “the machine stops” to better understand rational assumptions in algorithmic decision making and its implications for organizations. ***Academy of Management Review***, 45: 247-263.

Lipton, Z. C. 2018. The mythos of model interpretability. ***Queue***, 16: 31-57.

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. ***Proceedings of the 54th Hawaii International Conference on System Sciences***, 5463-5472.

Lumineau, F., Schilke, O., & Wang, W. 2020. Organizational trust in the age of the fourth industrial revolution. ***Working Paper***, available at http://dx.doi.org/10.13140/RG.2.2.20789.50401

Luo, X., Li, H., Zhang, J., & Shim, J. P. 2010. Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: An empirical study of mobile banking services. ***Decision Support Systems***, 49: 222-234.

Lyons, J. B., Aldin Hamdan, I., & Vo, T. Q. 2023. Explanations and trust: What happens to trust when a robot partner does something unexpected? ***Computers in Human Behavior***, 138, 107473.

Malhotra, D., & Murnighan, J. K. 2002. The effects of contracts on interpersonal trust. ***Administrative Science Quarterly***, 47: 534-559.

Mayer, K. J., & Argyres, N. S. 2004. Learning to contract: Evidence from the personal computer industry. ***Organization Science***, 15: 394-410.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. 1995. An integrative model of organizational trust. ***Academy of Management Review***, 20: 709-734.

McCulloch, W. S., & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. ***Bulletin of Mathematical Biophysic*****s**, 5: 115-133.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. 2011. Trust in a specific technology: An investigation of its components and measures. ***ACM Transactions on Management Information Systems***, 2: 1-25.

McKnight, D. H., Choudhury, V., & Kacmar, C. 2002. Developing and validating trust measures for e-commerce: An integrative typology. ***Information Systems Research***, 13: 334-359.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. ***Annual Review of Statistics and Its Application***, 8: 141-163.

Mori, M. 1970. The uncanny valley, ***Energy***, 7: 33-35.

Morris, M. W., Menon, T., & Ames, D. R. 2003. Culturally conferred conceptions of agency: A key to social perception of persons, groups, and other actors. ***Personality and Social Psychology Review***, 5: 169-182.

Murray, A., Rhymer, J., & Sirmon D. G. 2020. Humans and Technology: Forms of Conjoined Agency in Organizations. ***Academy of Management Review***, 46: 552-571.

Nass, C. 2004. Etiquette equality: exhibitions and expectations of computer politeness. ***Communications of the ACM***, 47: 35-37.

Nass, C., & Moon, Y. 2000. Machines and mindlessness: Social responses to computers. ***Journal of Social Issues***, 56: 81-103.

Newell, A., & Simon, H. A. 1976. Computer science as empirical inquiry: Symbols and search. ***Communications of the ACM***, 19: 113-126.

Nielsen, Y. A., Pfattheicher, S., & Keijsers, M. 2022. Prosocial behavior toward machines. ***Current Opinion in Psychology***, 43: 260-265.

Nienaber, A. M., & Schewe, G. 2014. Enhancing trust or reducing perceived risk, what matters more when launching a new product? ***International Journal of Innovation Management***, *18*(01), 1450005.

O'Connor, T. 1995. ***Agents, Causes, and Events: Essays on Indeterminism and Free Will***. New York, NY: Oxford University Press.

OECD. 2022. ***Recommendation of the Council on Artificial Intelligence***, OECD/LEGAL/0449.

Oneto, L., & Chiappa, S. 2020. Fairness in machine learning. In Oneto, L., Navarin, N., Sperduti, A., Anguita, D. (Eds.), ***Recent Trends in Learning From Data***. Springer International Publishing: Cham, Switzerland.

Parasuraman, R., & Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. ***Human Factors***, 39: 230-253.

Parkhe, A. 1993. Strategic alliance structuring: A game theoretic and transaction cost examination of interfirm cooperation. ***Academy of Management Journal***, 36: 794-829.

Pelau, C., Dabija, D. C., & Ene, I. 2021. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. ***Computers in Human Behavior***, 122: 106855.

Pirson, M., & Malhotra, D. 2011. Foundations of organizational trust: What matters to different stakeholders? ***Organization Science***, 22: 1087-1104.

Poppo, L., & Zenger, T. 2002. Do formal contracts and relational governance function as substitutes or complements? ***Strategic Management Journal***, 23: 707-725.

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. 2007. Comparing a computer agent with a humanoid robot. ***Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction***, 145-152.

Premack, D., & Woodruff, G. 1978. Does the chimpanzee have a theory of mind? ***Behavioral and Brain Sciences***, 1: 515-526.

Puranam, P., & Vanneste, B. S. 2009. Trust and governance: Untangling a tangled web. ***Academy of Management Review,*** 34: 11-31.

Rabin, M. 1993. Incorporating fairness into game theory and economics. ***American Economic Review***, 83: 1281-1302.

Raisch, S., Krakowski, S. 2020. Artificial intelligence and management: The automation-augmentation paradox. ***Academy of Management Review***, 46: 192-210.

Rindova, V. P., Williamson, I. O., Petkova, A. P., & Sever, J. M. 2005. Being good or being known: An empirical examination of the dimensions, antecedents, and consequences of organizational reputation. ***Academy of Management Journal***, 48: 1033-1049.

Ring, P. S., & Van de Ven, A. H. 1992. Structuring cooperative relationships between organizations. ***Strategic Management Journal***, 13: 483-498.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. 1998. Not so different after all: A cross-discipline view of trust. ***Academy of Management Review***, 23: 393-404.

Rosenthal-Von Der Pütten, A. M., & Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. ***Computers in Human Behavior***, 36: 422-439.

De Ruyter, B., Saini, P., Markopoulos, P., & Van Breemen, A. 2005. Assessing the effects of building social intelligence in a robotic interface for the home. ***Interacting with Computers***, 17: 522-541.

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. ***IBM Journal of Research and Development***. 3: 210-229.

Shapiro, D. L., Sheppard, B. H., & Cheraskin, L. 1992. Business on a handshake. ***Negotiation Journal***, 8: 365-377.

Shariff, A. F., Greene, J. D., Karremans, J. C., Luguri, J. B., Clark, C. J., Schooler, J. W., Baumeister R.F., & Vohs, K. D. 2014. Free will and punishment: A mechanistic view of human nature reduces retribution. ***Psychological Science***, 25: 1563-1570.

Shin, D., Zhong, B., & Biocca, F. A. 2020. Beyond user experience: What constitutes algorithmic experiences? ***International Journal of Information Management***, 52: 102061.

Sinha, R., & Swearingen, K. 2002. The role of transparency in recommender systems. In ***Proceedings of the CHI'02 Conference on Human Factors in Computing Systems***, 830-831.

Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. 2019. Organizational decision-making structures in the age of artificial intelligence. ***California Management Review***, 61: 66-83.

Stanovich, K. E., & West, R. F. 2000. Advancing the rationality debate. ***Behavioral and Brain Sciences***, 23: 701-717.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. ***Big Data & Society***, 9: 1-16.

Stea, D., Foss, K., & Foss, N. J. 2015. A neglected role for organizational design: Supporting the credibility of delegation in organizations. ***Journal of Organization Design***, 4: 3-17.

Strickland, L. H. 1958. Surveillance and trust. ***Journal of Personality***, 26: 200-215.

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., & Reynolds, N. 2019. Human factors challenges for the safe use of artificial intelligence in patient care. ***BMJ Health & Care Informatics***, 26: e100081.

Taddeo, M., & Floridi, L. 2018. How AI can be a force for good. ***Science***, 361: 751-752.

Talbert, M. 2019. Moral Responsibility in Zalta, E.N. (2019), ***The Stanford Encyclopedia of Philosophy*** (Winter 2019 Edition),
https://plato.stanford.edu/archives/win2019/entries/moral-responsibility/

Taylor, C. 1985. ***Human Agency and Language***. Cambridge, U.K.: Cambridge University Press.

Tjoa, E., & Guan, C. 2020. A survey on explainable artificial intelligence (XAI): Toward medical XAI. ***IEEE Transactions on Neural Networks and Learning Systems***, 32: 4793-4813.

Turing, A. M. 1950. Computing machinery and intelligence, ***Mind***, 59: 433-460.

Ullman, D., & Malle, B. F. 2018. What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In ***Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction***, 263-264.

Van der Woerdt, S., & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. ***New Ideas in Psychology***, 54: 93-100.

Vance, A., Elie-Dit-Cosaque, C., & Straub, D. W. 2008. Examining trust in information technology artifacts: The effects of system quality and culture. ***Journal of Management Information Systems***, 24: 73-100.

Vanneste, B. S. 2016. From interpersonal to interorganisational trust: The role of indirect reciprocity. ***Journal of Trust Research***, 6: 7-36.

Vanneste, B. S., Puranam, P., & Kretschmer, T. 2014. Trust over time in exchange relationships: Meta‐analysis and theory. ***Strategic Management Journal***, 35: 1891-1902.

Vanneste, B. S., & Gulati, R. 2022. Generalized trust, external sourcing, and firm performance in economic downturns. ***Organization Science***, 33: 1599-1619.

Venkatesh, V., & Davis, F. D. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. ***Management Science***, 46: 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. 2003. User acceptance of information technology: Toward a unified view. ***MIS Quarterly***, 27: 425-478.

Verberne, F. M., Ham, J., & Midden, C. J. 2015. Trusting a virtual driver that looks, acts, and thinks like you. ***Human Factors***, 57: 895-909.

Vergne, J. P. 2020. Decentralized vs. distributed organization: Blockchain, machine learning and the future of the digital platform. ***Organization Theory***, 1: 1-16.

Wang, W. 2017. Smartphones as social actors? Social dispositional factors in assessing anthropomorphism. ***Computers in Human Behavior***, 68: 334-344.

Wason, P. C., & Evans, J. S. B. 1974. Dual processes in reasoning? ***Cognition***, 3: 141-154.

Waung, M., McAuslan, P., & Lakshmanan, S. 2021. Trust and intention to use autonomous vehicles: Manufacturer focus and passenger control. ***Transportation Research Part F: Traffic Psychology and Behaviour***, 80: 328-340.

Waytz, A., Heafner, J., & Epley, N. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. ***Journal of Experimental Social Psychology***, 52: 113-117.

Weber, L., & Bauman, C. W. 2019. The cognitive and behavioral impact of promotion and prevention contracts on trust in repeated exchanges. ***Academy of Management Journal***, 62: 361-382.

Westerman, D., Edwards, A. P., Edwards, C., Luo, Z., & Spence, P. R. 2020. I-It, I-Thou, I-Robot: The perceived humanness of AI in human-machine communication. ***Communication Studies***, 71: 393-408.

Williams, M. 2001. In whom we trust: Group membership as an affective context for trust development. ***Academy of Management Review***, 26: 377-396.

Williams, M. D., Rana, N. P., & Dwivedi, Y. K. 2015. The unified theory of acceptance and use of technology (UTAUT): A literature review. ***Journal of Enterprise Information Management***, 28: 443-488.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. 2018. A qualitative exploration of perceptions of algorithmic fairness. In ***Proceedings of the CHI 2018 Conference on Human Factors in Computing Systems***, 1-14.

Wu, K., Zhao, Y., Zhu, Q., Tan, X., & Zheng, H. 2011. A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. ***International Journal of Information Management***, 31: 572-581.

Xu, K., Chen, X., & Huang, L. 2022. Deep mind in social responses to technologies: A new approach to explaining the Computers are Social Actors phenomena. ***Computers in Human Behavior***, 134: 107321.

Yamagishi, T., & Yamagishi. M. 1994. Trust and commitment in the United States and Japan. ***Motivation and Emotion***, 18: 129-166.

Zaheer, A., McEvily, B., & Perrone, V. 1998. Does trust matter? Exploring the effects of interorganizational and interpersonal trust on performance. ***Organization Science***, 9: 141-159.

Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. 2021. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers, ***Working Paper***, available at https://arxiv.org/abs/2105.02117

Zhang, L., & Yencha, C. 2022. Examining perceptions towards hiring algorithms. ***Technology in Society***, 68: 101848.

TABLE 1
Predictions from the Model of Perceived Agency and Trust in AI

Implication for	Contingency	Prediction
Static trust in AI	Task complexity	When the task is more complex, greater agentic perceptions increase human trust in AI by enhancing ability perceptions.
	Designer reputation	When the designer's reputation is poor, greater agentic perceptions increase human trust in AI by shifting attention from the designer to the AI.
	Potential loss	When the potential loss from AI usage is high, greater agentic perceptions decrease human trust in AI because of the salience of betrayal aversion.
Dynamic trust in AI	Application novelty	When the AI application is novel, greater agentic perceptions will lead to greater trust increases in the AI after successful interactions because humans begin with lower expectations about it.
	Trust breach	After a trust breach by the AI, greater agentic perceptions will lead to a greater decrease in trust in AI because it is seen as more responsible for its behavior.
	Designer scandal	When the designer is involved in an unrelated scandal, greater agentic perceptions lead to a lower reduction in trust in AI because perceived trustworthiness of the designer matters less.
Dynamic agency perceptions of AI	Explainability	The greater the capacity of AI to explain its decisions, the less will be the perceived agency of the AI because of reduced opacity.
	Updating	The greater the capacity of the AI to update its model with data from new cases, the greater will be the perceived agency of the AI because of its ability to learn.

TABLE 2
Interventions that Change Trust in AI

Changing perceptions of	Interventions	Examples	Conditions under which interventions increase trust in AI
AI's agency	AI's decision autonomy	Let a robot decide on the course of action (Kim & Hinds 2006)	Increase AI's perceived agency if AI's perceived trustworthiness is high or if designer's perceived trustworthiness is low Decrease AI's perceived agency if AI's perceived trustworthiness is low or if designer's perceived trustworthiness is high
	Anthropomorphizing	Give a name, gender, and human voice to an autonomous vehicle (Waytz et al. 2014)	
	Transparency (reveal how algorithm works)	Show the model that underlies an AI	
AI's trustworthiness	Reliability	Enhance AI's performance on a test set	Increase AI's perceived trustworthiness if AI's perceived agency is high or if designer's perceived trustworthiness is low
	Explainability (provide a justification for specific actions)	Let robot explain the reason for its action (Kim & Hinds 2006)	
Designer's trustworthiness	Communication	Designer commits to an AI safety standard	Increase designer's perceived trustworthiness if AI's perceived agency is low or if AI's perceived trustworthiness is low
	Reputation	Form an alliance with a well-respected partner (Higgins & Gulati, 2003; Rindova et al., 2005)	

FIGURE 1
Trust Situations with Non-agentic and Agentic Technologies

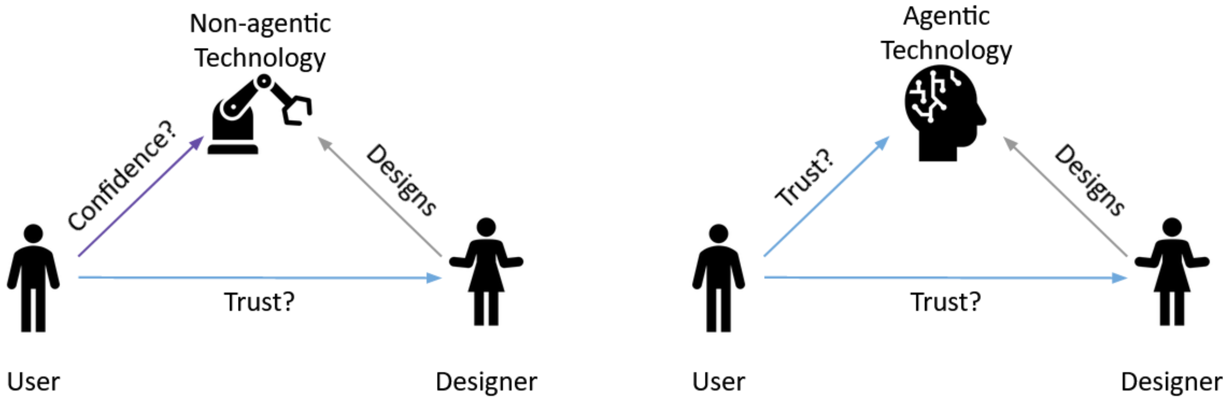
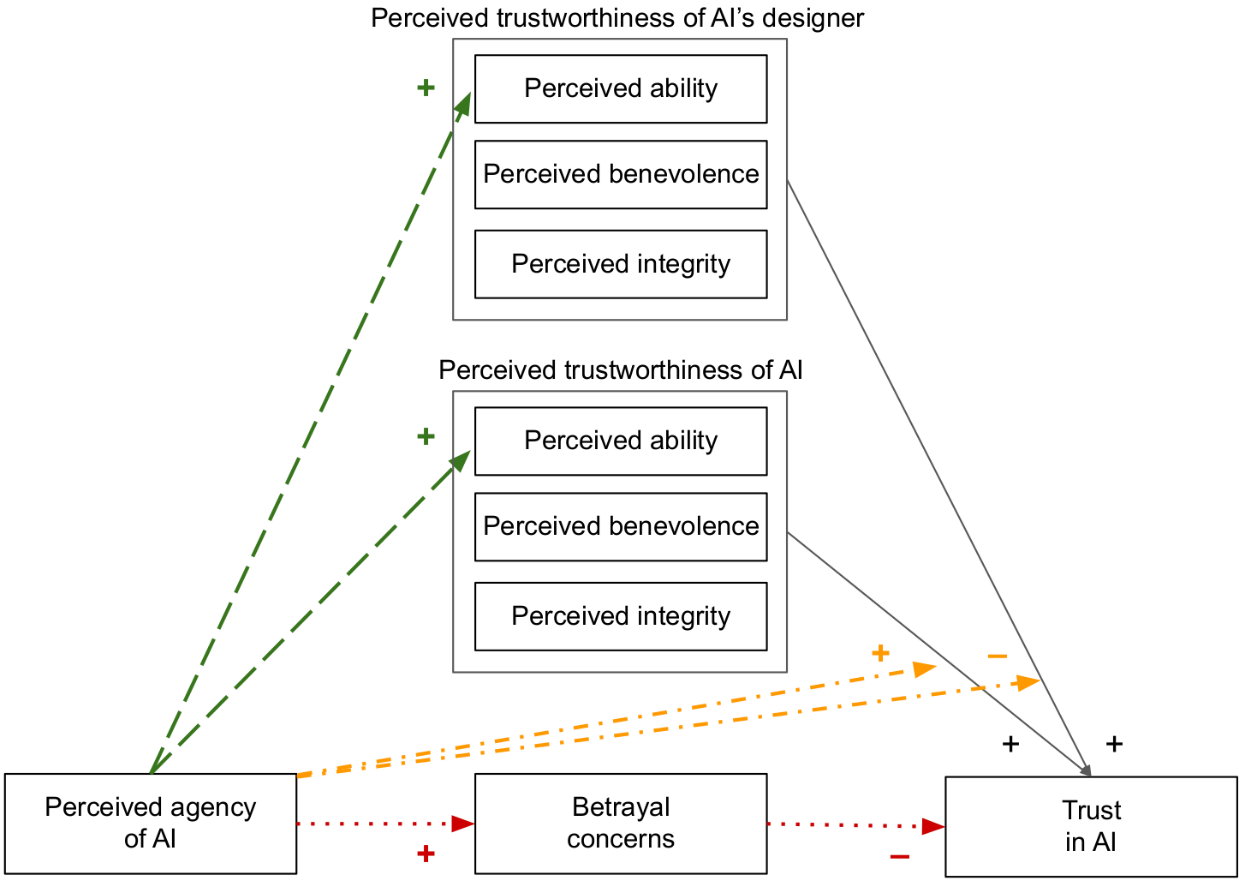


FIGURE 2
A Model of Perceived Agency and Trust in AI



Notes: Full, black relationships capture relationships established in the existing literature. Dashed, green relationships represent enhancing perceived ability; dash-dotted, orange relationships represent shifting the locus of trust (higher perceived agency increases the importance of the perceived ability, benevolence, and integrity of the AI (+) and decreases the importance of these of the AI's designer (-)); and the dotted, red relationships represent amplifying betrayal concerns.