Working Paper

2021/42/STR

Artificial Intelligence, Trust, and Perceptions of Agency

Phanish Puranam
INSEAD, Phanish.puranam@insead.edu

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

30 July 2021

Extant theories of trust assume the trustee has agency (i.e. intentionality and free will). We propose that a crucial qualitative distinction between placing trust in Artificial Intelligence (AI) vs. trust in a human lies in the degree to which attributions of agency are made to the trustee by the trustor (human). We specify two mechanisms through which the extent of agency attributions can affect human trust in AI. First, the importance of the benevolence of the trustee—the AI—increases if the AI is seen as more agentic, but so does the anticipated psychological cost if it violates the trust (because of betrayal aversion, see Bohnet & Zeckhauser, 2004). Second, attributions of benevolence and competence become less relevant for placing confidence in a non-agentic seeming AI system, and instead benevolence and competence attributions to the designer of the system become important. Both mechanisms imply that making an AI appear more agentic may increase or decrease the trust that humans place in it. While designers of AI technology often strive to endow their creations with features that convey its benevolent nature (e.g. through anthropomorphizing or transparency), this may also change agency perceptions in a manner that results in making it less trustworthy in human eyes.

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

Acknowledgement: Puranam thanks the Desmairis Fund for research support for the Organizations and Algorithms project.

INTRODUCTION

Artificial Intelligence (AI) is seen today as a transformative technology which has created the possibility of large scale human reliance and possibly dependence on non-human intelligences. The number of arenas in which humans are coming to interact with and rely on AI's has increased dramatically in the recent past, raising the importance of the existence of trust (or lack thereof) between humans and AI. An emerging stream of literature has investigated the nature of trust by humans in AI, including robots. A variety of factors that affect human trust in AI have been documented, sometimes with conflicting results.

However, this literature has largely carried over the premises about how trust operates between humans to an understanding when and how humans trust AI. In particular, a crucial distinction between trust among humans vs. between humans and AI has remained, in our view, under-appreciated: namely the extent to which perceptions of agency about the trustee by the trustor shape trust. To illustrate this distinction, consider two questions: a) what is the difference between trusting an Artificial Intelligence (AI) vs. a human investment manager to make an investment recommendation? b) What is the difference between trusting an excel spreadsheet to identify the investment with the highest expected returns and an AI that recommends the investment?

The purpose of this paper is to argue that an important qualitative distinction between these situations lies in the attributions of agency (i.e. intentionality and free will) to the trustee by the trustor (human). As classic definitions of trust have made clear, trust is only defined in the context of an agentic attribution. Absent such an attribution, the problem is not one of trust but of confidence in inanimate technology. This is what separates the excel spreadsheet from the AI advisor in the second question. Theories of trust developed so far assume as a boundary condition that the trustee has agency. However, AI, particularly machine learning driven AI occupies a curious place somewhere between humans and technology in the extent to which it is attributed agency. Since it is often embodied in a goal directed learning system, it may appear to some degree to possess both intentionality (goal directed behavior) and free will (the ability to change and modify immediate objectives), thus appearing agentic. Older AI systems that were rule based in contrast may have appeared intentional but were less likely to have conveyed the appearance of possessing free will. This produces attributions of agency to AI that may be in excess of that for a technology like excel, but possibly less than what one routinely attributes to fellow humans. This is what makes the answer to the first question problematic.

This paper will pursue a systematic answer to questions similar to the first one as follows:. building on past conceptualizations of the antecedents of trusting behaviour, we specify two mechanisms through which the degree of agency attributions can affect human trust in AI. The first mechanism depends on a tension between benevolence and betrayal aversion. The importance of the benevolence of the trustee—the AI—increases if the AI is seen as more agentic, but so does the anticipated psychological cost if it violates the trust (because of the well documented phenomenon of betrayal aversion, see Bohnet & Zeckhauser, 2004). The second mechanism involves a shift in the locus of trust from the AI to its designer as perceptions of the AI's agency decline. Attributions of benevolence become less relevant for placing confidence in a non-agentic seeming AI system, conditional on benevolent attributions to the designer of the system.

Both mechanisms imply that making an AI appear more agentic may increase *or* decrease the trust that humans place in it in a systematic manner. While designers of AI technology often strive to endow their creations with features that convey its benevolent nature (e.g. through anthropomorphizing or transparency), this may also change agency perceptions and so make it less trustworthy in human eyes. Our analysis suggests that it may be possible to increase human willingness to rely on AI by *reducing* their features that convey agency, if credibly signalling benevolence is difficult, and instead enhancing signals of benevolence and competence of the designer.

In sum, we propose that taking this into account the role of perceptions of agency will help us better integrate and explain some of the current findings about when humans trust AI, as well as make new falsifiable predictions for testing in future studies. Further we also hope to contribute to the general study of trust, which has noted the importance of high levels of agency attributions for trust, but not the impact of variations in the degree of the perception of agency.

THE IMPORTANCE OF AGENCY PERCEPTIONS FOR TRUST

Trust requires a perception of agency

Trust involves positive attributions by the trustor about the competence and benevolence/integrity of the trustee. Attributions about benevolence matter independently of attributions of competence (i.e. expertise or technical adequacy)—we do not trust a well-meaning bumbler. We will here focus on this benevolence aspect since it is a necessary ingredient to trust.

An often used definition of trust that emphasizes the benevolence aspect is "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" (Mayer et al., 1995: 712). This definition makes clear trust does not apply without the trustee having agency.

Vulnerability is defined relative to another party that has agency. For example, playing a lottery against nature, even if risky, or placing confidence in your car not breaking down are not instances of trust (though colloquially, the word may be applied). Trust is relevant only because the trustee has some choice between acting in a way that may harm or benefit the trustor (i..e agency). For example, extensive governance mechanisms limits trust development 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).

Hardin (2002) captures the centrality of agency attributions to trust very well. He notes

> *"[...] one might suppose it perverse to say, "I trust you to do Y," when it is in your interest to do Y. 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.*

An additional implication is that the lack of (or low *levels* of) agency attributions can impede trust building. While agency may be present or absent, its perceptions need not be dichotomous. For instance, in exchange relationships that occur in the context of binding contracts (Malhotra & Murninghan,2002), the development of trust over time may be impeded because there is limited scope for the partners to potentially engage in opportunistic behavior (though not zero scope since all contracts are incomplete). This is what Puranam & Vanneste (2009) call the indirect crowding out effect of contracts on trust—distinct from the direct effect of imposing a contract and thus signalling distrust. In sum, perceptions of agency are necessary for trust to exist, and low levels of agency perceptions may impede the formation of trust.

Perception of agency rest on beliefs about intentionality and free will

Perceptions of agency in turn require two things. First it requires an attribution of "intentionality"—the capacity to understand means-ends relationships and make decisions based on these (same as decision making capability or ability to think/plan and act) (Gray et al., 2007). In interactions with fellow humans we make this attribution routinely, though possibly to varying degrees, adjusting for incapacity (e.g. being unhealthy or a minor- as is embodied in the legal formula of being "an adult of sound mind and body").

Second, it also requires the attribution of “free will”—the ability to choose one's own goals. 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 very useful entries on the topic in the Stanford Encyclopedia of Philosophy[1]). Instead, we highlight the aspects of free will that we believe are central to our current concerns with human trust in AI.

Informally, free will can be seen as the power to control one’s choices and actions. Subjectively, the experience of free will is the perception of volition, choice and the ability to imagine that we might have chosen otherwise (Baumeister et al., 2008). The attribution of free will involves assuming that a party one is interacting with also has this property.

There are strong arguments against the existence of free will. The principle of causal determinism which can be found in the work of Spinoza, Liebniz and Descartes, and which is foundational to most modern fields of scientific inquiry, assumes that every event is determined by a prior chain of causes that operate according to the laws of nature. This implies that there are no uncaused causes and so rules out forms of free will that assume such an ontological status for individuals (Clarke, 1996; O’Connor, 2005).

An important direction that discourse on free will has taken has been one of identifying what form of free will is necessary to assign moral responsibility to actors. The perspective known as “compatibilism” suggests that causal determinism can be compatible with moral responsibility, though it is critiqued by causal determinists who argue that it makes little sense to hold an actor morally responsible if that actor does not have freedom to choose their course of action. A complication here is the distinction between freedom to choose conditional on goals, vs. the freedom to choose the goals themselves. The first is a

[1] https://plato.stanford.edu/

matter of external constraints, and one may be more (or less) free of them. Compatibilists use this argument to support the assignment of moral responsibility. The second seems to admit less freedom given the principle of causal determinism cited above, and causal determinists use this argument for moral exculpability.

Regardless of whether we actually have free will, the belief that we do (and the imputation of others having it) both seem to play an important role in social behavior (Wertenbroch et al, 2008). As many have suggested (e.g. Kane 1996), such a belief may be essential to give meaning to constructs such as striving for achievement, autonomy, dignity, love, and friendship. We suggest trust is no exception. The belief in the free will of the trustee is essential for the trustor to impute agency to them.

In Table 1, we decompose perceptions of agency into three hierarchical levels based on its component attributions: intentionality and free will. The discrete decomposition into levels is for expositional clarity: perception of intentionality and free will may vary continuously in practice. We consider cases when neither are present, intentionality attributions are present (which is necessary for a perception of free will) and a perception of free will is also present. The first key implication of this decomposition is that perception of agency about a system—whether a human or AI—can vary based on the component attributions it is composed of.

[[INSERT TABLE 1 ABOUT HERE]]

Attributions of benevolence only become relevant given perception of agency, Benevolence is only defined for an entity with agency—the ability to act autonomously in a goal oriented manner. Attributions of agency are necessary but not sufficient for attributions of benevolence. An agent to be seen as benevolent must be seen as likely to take favorable actions to oneself despite facing no compulsion to do so, except benign intentions themselves. The two aspects can exist without the other. An enemy warrior is an agent but not benevolent, a house cleaning robot is benevolent only to the extent it is seen as an agent, else it is simply beneficial (or munificent). It is only when both agency and benign motivations are attributed that benevolence and therefore trust truly exists. The necessity of agency to assign benevolence is recognized by philosophers of ethics who often use the term moral agency synonymously with agency, since one cannot attribute moral responsibility without agency (e.g. Gray et al., 2007; Bigman et al., 2019).

In fact, the attributions of agency may raise doubts about benevolence. Specifically, attributions of agency may give rise to concerns about betrayal that do not exist when interacting with a non-agent. In a series of studies, Bohnet and colleagues (Bohnet & Zeckhauser, 2004; Bohnet et al., 2008, 2010) have shown that when interacting with an agent (as opposed to making a risky decision), we experience concerns not only about fairness in terms of relative allocations but also about the possibility of being betrayed. The same negative outcome, when resulting from betrayal produces more disutility than if it was the result of ill-luck, and this disutility cannot be accounted for in terms of fairness preferences (i.e. how the outcome compares to that received by others) alone. A second implication of the separation into three levels of perception of agency therefore is that benevolence attributions matter only at some of them (see last column).

This decomposition of attributions of benevolence into attributions about agency + attributions of benign motives is not normally necessary when dealing with either humans, because agency is presumed (focusing the action on benign motivations alone), or machines, because it is presumed absent (making trust irrelevant). But with AI, particularly the kind that seems to act autonomously through learning, there may be degrees to which attributions of agency are made—the perception of agency may lie somewhere between Level 1 and Level 2. This, we argue, plays a large role in whether humans trust AI and what we can do about it (also see Bigman et al. (2019) on judgements of robot responsibility).

A MODEL OF PERCEIVED AGENCY AND TRUST

Our proposed model can be summarized in Figure 1. We focus for now on an interaction between a human, an AI system and its designer. Here, trust, competence, benevolence, and agency perceptions retain their meanings as defined in previous sections. Based on prior research (Mayer et al., 1995), the baseline model is that perceptions of benevolence and competence enhance trust in AI. We add that perceived agency may strengthen or weaken these relationships (an indirect effect) and that perceived agency may directly affect trust in AI (a direct effect).

[[INSERT FIGURE 1 ABOUT HERE]]

Indirect effect

The first key idea in Figure 1 is to organize the relationships between the perceptions of benevolence and competence and trust in AI 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 attributions of benevolence and competence made to the system *or* to the designer of the system. At the extremes, the relevant attributions are only to the system itself if significant levels of agency are attributed to it. Else, the attribution is entirely to the system's designers. More generally both sets of attributions may operate simultaneously, but from the human's point of view, whether the interaction is with the AI or with it's designer is tuned by the attribution of agency the human makes to the AI. One might say that there is always a triad of entities in this interaction, but which dyad is salient to the human depends on agency perceptions about AI.

The attribution of agency itself is the result of two factors—attributions of intentionality and attributions of free will to the system. A system is treated as intentional if it appears to be a decision making entity—it can alter behavior to suit a context. Note that intentionality is not the same as competence. A system can be competent ("fit for purpose") without being intentional, and a system can display intentionality without being competent. Flexible behavior in response to contextual cues is key to judging intentionality. An attribution of intentionality is necessary but not sufficient for an attribution of free will, as discussed in the previous section.

There are two main implications of the theoretical relationships set out in Figure 1. First, benevolence and competence are functional equivalents; trust can be increased either by increasing perceptions of benevolence or competence, and if trusting behavior requires a threshold level of trust, then these can also serve as functional substitutes (Puranam & Vanneste 2009). Second, attributions of competence and benevolence matter the most for trust in AI when it appears to display intentionality and free will. This implies that increasing agentic behavior of an AI may lead to decreased rather increased trust in it by humans unless an increase in benevolence attributions simultaneously can be produced.

A direct effect

The second key idea in Figure 1 is that perceived agency of AI may directly affect trust in AI. For a human trustee, Bohnet and colleagues (Bohnet & Zeckhauser, 2004; Bohnet et al., 2008, 2010) have established that the anticipated psychological costs of a 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 betrayal but not the associated psychological costs of being betrayed. As we argued above, AI is in between nature and a human in terms of perceptions of it's agency (between Level 1 and Level 2). Enhancing the perceived agency of AI gives 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 reduces trust in the AI because of the additional concern that betrayal would be painful.

Their joint effect

We discussed the indirect and direct effect separately but they should be analyzed jointly if the goal is to understand when human trust in AI is likely to increase (or decrease). For example, an alternative to increasing attributions of either AI's benevolence or competence is to decrease the attributions of agency—to shift the emphasis instead to attributes of benevolence and competence of the designer of the system. Thus, if it is difficult to improve attributions of benevolence towards an AI, it may be better to de-emphasizes its agentic attributes and emphasize instead the benevolence of its designers. These implications help to analyse a range of possible interventions to increase human trust in AI, with some predictions on when they are most and least likely to succeed.

APPLICATIONS: TRUST INTERVENTIONS

Practitioners and academics are interested in trust in AI because a lack thereof may prevent the adoption or use of a beneficial AI system. The conceptual model allows us to anticipate the effect of interventions on trust in AI. We discuss four such interventions (see Table 2).

[[INSERT TABLE 2 ABOUT HERE]]

The first intervention is ***autonomy*** or giving AI a greater ability to decide on the course of action (Kim & Hinds, 2006; Waytz et al., 2014). As a result, agency perception of AI should increase as it is deemed to have greater intentionality and free will (Kim & Hinds, 2006). We do not anticipate an effect on benevolence perceptions, as increasing autonomy does not necessarily mean that an AI is more likely or not to do what is good for the human. The competence perception may also increase if an AI system that is autonomous is seen as more likely to execute well the task at hand. In fact, trust will only increase if more autonomy comes with higher perceived competence (because higher perceived agency triggers betrayal aversion concerns). Else, an increase in autonomy may lower trust.

The second intervention is ***reliability*** or increasing the performance of the AI system. This intervention should enhance competence perceptions of AI but not necessarily affect benevolence or agency perceptions. Taken together, trust in AI is expected to go up, which is consistent with the current empirical evidence (Glikson & Woolley, 2020).

The third intervention is ***transparency*** or explaining the logic for AI's decisions or actions. Many AI algorithms are black boxes in that a human does not see the reasons for a decision. The lack of transparency is common when AI relies on complex functional forms, including neural networks. Efforts are increasingly made to provide explanations for AI's decision making or so called explainable AI (XAI). Benevolence perceptions are unaffected if the explanation of a decision does not make the decision seem more favorable to the human. Competence perceptions are enhanced if decisions are seen as less random. Bigman et al. (2019) argues that opacity increases perceptions of free will by reducing predictability. Kim & Hinds (2006) argued that transparency makes a robot less likely to be blamed. Making AI more explainable may thus detract from its capacity to be treated as an agent. This is analogous to the sense of diminished moral responsibility we ascribe to humans as we understand more about how their minds work. Overall, trust in AI should increase with transparency because betrayal aversion concerns are lessened.

The fourth intervention is ***anthropomorphizing*** or making AI more human-like, not merely in terms of physical appearance but also in terms of the attribution of human mind-like qualities. 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 it is seen as having a mind of its own. To the extent that through anthropomorphizing the AI comes to more resemble a human, then benevolence perceptions will increase (Glikson & Woolley, 2020). One mechanism is that of increased homophily, whereby others like us are trusted more, or relatedly, of decreased speciesism, whereby other species are trusted less. The effect on competence perceptions could go either way. If anthropomorphizing leads to competence perception of AI similar to those of a human, then competence perception of AI may increase or decrease depending on whether a human was seen as more or less competent. 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.

Additional interventions include decreasing agency perceptions of AI and increasing benevolence or competence perceptions of AI's designer (Eslami et al., 2015; Hengstler et al., 2016).

CONCLUSION

Existing theories of trust assume as a boundary condition that the trustee has agency. 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. We use the case of human trust in AI to analyze the role of variable degrees of agency perception on trust, and the broader implications for a new understanding of trust among humans. 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 vs. owner of a business.

It has long been recognized that uncertainty is crucial for trust. As Hardin (2002) noted:

> *"Many writers therefore suppose that trust is inherently embedded in uncertainty. 'For trust to be relevant,' Diego Gambetta (1988, 218-19) says, 'there must be the possibility of exit, betrayal, defection,' by the trusted (see also Yamagishi and Yamagishi 1994, 133; and Luhmann 1979). More generally, 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. Trust as well as choice and rationality are at issue just because we are in the murky in-between land that is neither deterministic nor fully indeterminate."*

Our approach clarifies that while uncertainty and agency perceptions are both necessary for the phenomenon of trust (as defined) to exist, there can be degrees of not only uncertainty but also agency perception. Further, these may be independent of each other in some contexts (e.g. high uncertainty combined with low agency perceptions). Working out the theoretical and empirical implications of these differences seems a fruitful avenue for expanding our understanding of trust.

REFERENCES

Baumeister, R. F., Sparks, E. A., Stillman, T. F., & Vohs, K. D. (2008). Free will in consumer behavior: Self‐control, ego depletion, and choice. *Journal of Consumer Psychology*, *18*(1), 4-13.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365-368.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, *55*(4), 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*(1), 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*(2), 811-828.

Clarke, R. (1995). Indeterminism and control. *American Philosophical Quarterly*, 32, 125−138.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... & Sandvig, C. (2015, April). "I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 153-162).

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. science, 315(5812), 619-619.

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

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.

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

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

Malhotra, D., & Murnighan, J. K. (2002). The effects of contracts on interpersonal trust. *Administrative Science Quarterly*, *47*(3), 534-559.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734.

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

Puranam, P., & Vanneste, B. S. (2009). Trust and governance: Untangling a tangled web. *Academy of Management Review, 34*(1), 11-31.

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

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.

Wertenbroch, K., Vosgerau, J., & Bruyneel, S. D. (2008). Free will, temptation, and self‐control: We must believe in free will, we have no choice (Isaac B. Singer). *Journal of Consumer Psychology*, *18*(1), 27-33.

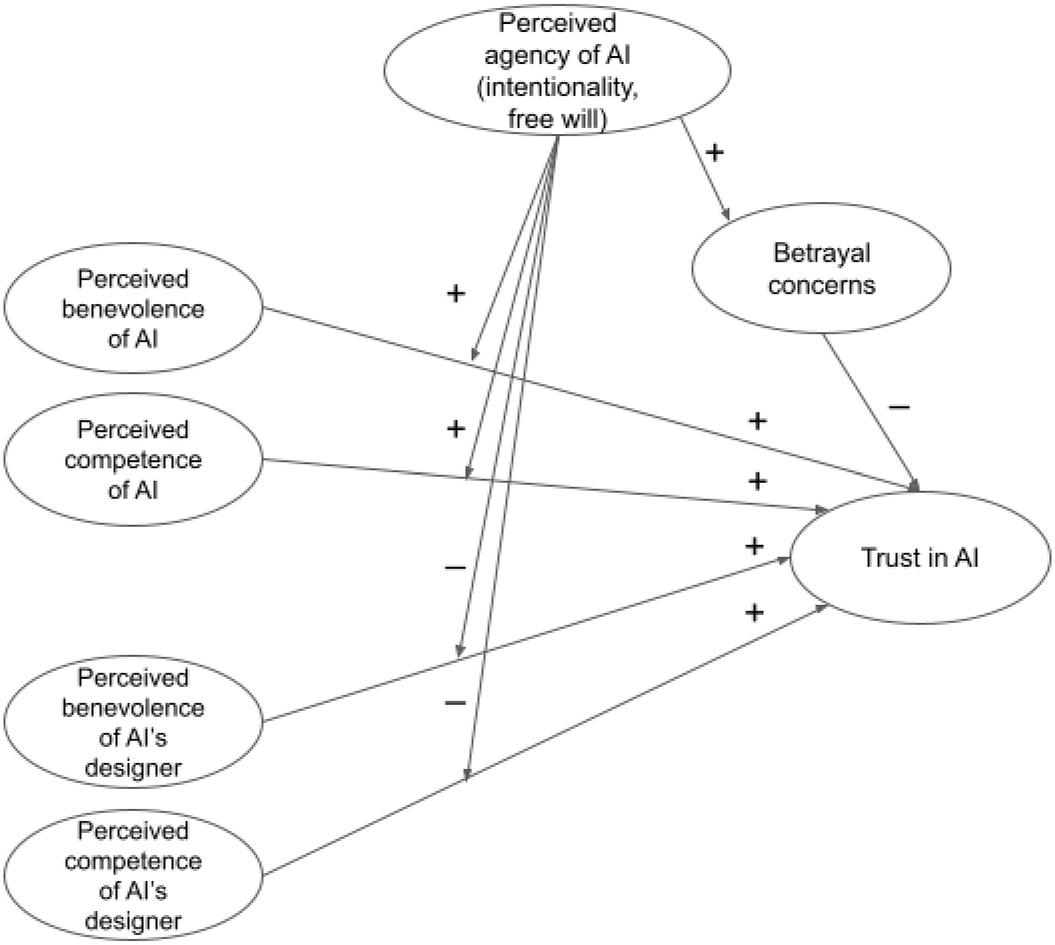
TABLE 1. LEVELS OF AGENCY ATTRIBUTIONS

	Attributions of Intentionality	**Attributions of Free will**	**The relevance of attributions about motivations**
Level 2 Human (and possibly AI based on ML)	Yes	Yes	Attributions about motivations of alter matter because attributions of agency triggers concerns about betrayal aversion (agentic harm to self) and perhaps also social preferences (fairness)
Level 1 (e.g. Old rule based AI)	Yes	No	The agent itself has no motives, but the designer does.
Level 0 Nature, a lottery	No	No	The agent itself has no motives and there is no designer.

TABLE 2. TRUST INTERVENTIONS AND THE ROLE OF PERCEPTIONS OF AI

Intervention	**Effect on perceptions of AI of**			**Effect on Trust in AI**
	Competence	**Benevolence**	**Agency**	
Autonomy	Increase	None	Increase	Indeterminate
Reliability	Increase	None	None	Increase
Transparency	Increase	Increase	Decrease	Increase
Anthropomorphizing	On par with human	Increase	Increase	Indeterminate

FIGURE 1. A MODEL OF PERCEIVED AGENCY AND TRUST IN AI


Perceived agency of AI (intentionality, free will)
+
Betrayal concerns
Perceived benevolence of AI
+
+
–
Perceived competence of AI
+
+
+
Trust in AI
–
+
–
Perceived benevolence of AI's designer
Perceived competence of AI's designer