Annals of Operations Research (2022) 308:525–548
https://doi.org/10.1007/s10479-020-03702-9

S.I.: ARTIFICIAL INTELLIGENCE IN OPERATIONS MANAGEMENT

Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems

Yulia Sullivan[1] · **Marc de Bourmont**[2] · **Mary Dunaway**[3]

Published online: 17 July 2020

Abstract
As artificial intelligence (AI) becomes more pervasive, the concern over how users can trust artificial agents is more important than ever before. In this research, we seek to understand the trust formation between humans and artificial agents from the morality and uncanny theory perspective. We conducted three studies to carefully examine the effect of two moral foundations: perceptions of harm and perceptions of injustice, as well as reported wrongdoing on uncanniness and examine the effect of uncanniness on trust in artificial agents. In Study 1, we found perceived injustice was the primary determinant of uncanniness and uncanniness had a negative effect on trust. Studies 2 and 3 extended these findings using two different scenarios of wrongful acts involving an artificial agent. In addition to explaining the contribution of moral appraisals to the feeling of uncanny, the latter studies also uncover substantial contributions of both perceived harm and perceived injustice. The results provide a foundation for establishing trust in artificial agents and designing an AI system by instilling moral values in it.

Keywords Artificial intelligence · Perceptions of harm · Perceptions of injustice · Uncanniness · Trust

1 Introduction

In the last few years, scientists, researchers, and engineers have made a remarkable progress in machine learning and artificial intelligence (AI) systems. The Oxford English Dictionary defines an AI as "the theory and development of computer systems able to perform tasks normally requiring human intelligence" (Abrardi et al. 2019, p. 1). Recently, this topic is attracting the attention of researchers from the information systems (IS) and operational management (OM) fields (e.g., Beck and Smith 2009; Brucker and Knust 2002; Fragapane

✉ Yulia Sullivan
yulia_sullivan@baylor.edu

1 Baylor University, Waco, TX, USA

2 NEOMA Business School, Mont-Saint-Aignan, France

3 Morgan State University, Baltimore, MD, USA

et al. forthcoming; Han and Cook 1998; Kats and Levner 1997; Petrovic 2019; Seeber et al. forthcoming; Talbi 2016). In this current study, we define artificial intelligence as a system-based artificial agent (i.e., artificial agent)—a system that is capable of sensing, information processing, decision-making, and learning to act upon its environment and to interact with humans and other machines in order to achieve a common goal (Seeber et al. forthcoming). Indeed, artificial agents now can perform tasks that regularly require human interventions, such as diagnosing illnesses, making decisions on the factory floors, and even engaging in a conversation with humans. The adoption of artificial agents creates entirely new opportunities for flexible, efficient production, even when it comes to complex and increasingly customized products in small batch runs (Siemens 2019). According to PwC study, a total of 62% of large companies are already utilizing AI systems in 2018 and these systems can contribute to up to US$ 15.7 trillion to the global economy in 2030 (PwC 2017).

As artificial agents become more pervasive, the concern over how we can trust them is more important than ever before. While there are a lot of potentials for artificial agents to improve our lives, there is just as much potential for it to cause harm or inequality. Moreover, some fundamental questions arise about the expected social impact of artificial agents, especially if the interaction with an intelligent agent involves a wrongdoing act. What people find appropriate for a machine may be different than what people are going to accept from a human (IBM 2019). Given the fact that the mechanism of trust formation in the interaction between humans and artificial agents has not been fully studied, little is known about what factors contribute to trust in artificial agents. Trust is important in situations where there is a state of dependence between two parties and when this dependence entails risk (Komiak and Benbasat 2006). In the context of users' interaction with an artificial agent, given the amount of autonomy granted to the agent to achieve a shared goal, users depend on the agent for better decision making or task performance. Risk arises because decisions or actions performed by an artificial agent can have serious consequences for society at large, and even greater consequences for the companies responsible (Saif and Ammanath 2020). Thus, the adoption of AI systems is largely dependent on trust.

The goal of the current study is to understand the trust formation between humans and artificial agents by studying this phenomenon from the morality and uncanny theory perspective. Most of the literature on the relationship between humans and artificial agents has focused on whether a robot, as a subset of artificial agents, may look uncanny when human and nonhuman elements are mixed (e.g., Gray and Wegner 2012). Whereas these studies specifically focus on the aesthetic aspects of artificial agents to increase the likeability of a humanoid object, the possibility that an eerie feeling or uncanny feeling can be triggered by moral and personality aspects of the artificial agents has never been explored. If one's judgments of moral wrongdoing can trigger the uncanny feeling towards artificial agents, the important questions are "what types of moral appraisals will likely drive the uneasy feeling?" and "what is the impact of moral violation on trust between humans and artificial agents?".

The current research examines whether the uneasy feeling or uncanny feeling will mediate the effect of moral appraisals on trust in artificial agents. Trust is crucial because it impacts the success of human-nonhuman agent interactions and may determine the future adoption of AI systems (Piazza et al. 2019). Although prior research has identified the importance of human characteristics (e.g., ability, personality), environmental characteristics (e.g., task, team), and robot characteristics (e.g., performance, attributes) on human-artificial agents team trust (e.g., Billings et al. 2012), it did not take into account the social expectations about how artificial agents should act in a socially constructed environment. For example, is society in general willing to trust an artificial agent that demonstrates bias toward a specific social group? Whereas such bias is inevitable among human agents, people expect an artificial

agent to make a better and more objective decision. In a case of moral event as described in this example, we hypothesize that the uncanny feeling will be activated and in turn, it may influence observers' level of trust in nonhuman agents.

In this current study, we specifically focus on two types of moral appraisals: *harm appraisal* and *injustice appraisal* as the predictors of the uncanny feeling. According to the harm-centric approaches to moral wrongdoing, people judge an act as morally wrong if they perceive the act to cause harm (Piazza et al. 2019). From this theoretical lens, harm constitutes a foundational, organizing template by which all immoral actions are conceptualized (Gray and Schein 2012; Gray et al. 2014; Piazza et al. 2019). However, some scholars have argued that perceptions of harm are not sufficient for judgments of wrongdoing because people often find harmful acts acceptable (e.g., Piazza and Sousa 2016). Although there is limited research investigating how people react when a nonhuman agent causes harm to humans, studies on autonomous vehicles have shown that people tend to approve the vehicle's act to kill one to save five people (e.g., Awad et al. 2018; Bonnefon et al. 2016). Certainly, perceptions of harm are observed in these studies. However, it is not sufficient for judgments of wrongdoing. On the contrary to the harm-centric perspective, some have argued that if a harmful act is apprised as involving injustice, then it is judged to be morally wrong (Sousa and Piazza 2014). In this regard, the injustice appraisal is the appraisal that the agent did not consider the balance of interests involved when causing the pain/suffering (Piazza et al. 2019). We predict that, these two cognitively moral appraisals—harm appraisal and injustice appraisal—will trigger an emotional appraisal (i.e., the uncanny feeling) among observers since there is an intellectual uncertainty when an artificial agent's behavior violates expectations. This uncanny feeling, in turn, will negatively influence trust in artificial agents.

Our research sheds light on the social determinants of trust between humans and artificial agents from the moral judgment perspective. Whereas previous studies on trust between human and nonhuman objects have primarily focused on robots, our research focuses on a broader category of AI systems. Artificial agents can be conceptualized based on their degree of autonomous agency as information retrieval agents (e.g., commercial shopping chatbots), advisory agents (i.e., using machine learning to provide information about a defendant that can be used for a parole decision), and performative agents (e.g., scheduling and delivering) (Nissen and Sengupta 2006; Seeber et al. forthcoming). Artificial agents can also be categorized based on its level of embodiment as embodied agents (e.g., humanoid robot) and disembodied agents (e.g., chatbot, digital assistant) (Seeber et al. forthcoming). For remainder of this paper, we use the term *artificial agents* and *AI systems* interchangeably to refer to any of these types of AI, which implement at least one subset of these functions, but not necessarily all of them. In one survey and two situated experiments, we demonstrate the uncanny feeling mediates the relationships between moral appraisals and trust in artificial agents. Whereas in Study 1, we do not differentiate the agent functions, in Study 2 we focus on performative, embodied agents (i.e., manufacturing robot) and in Study 3 we focus on advisory, disembodied agents (i.e., machine-learning based recruitment system). The use of these different scenarios is intended to increase the robustness of our model.

2 Theory and hypotheses

2.1 Moral foundational theory (MFT)

According to MFT, the human mind is organized in advance of experience so that it is prepared to learn values, norms, and behaviors related to a diverse set of recurrent adaptive social

problems (Graham et al. 2013). Unlike other explicit deliberative reasoning and decision-making processes, moral judgments tend to happen quickly (Graham et al. 2013). This moral evaluation process is described as "the sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like-dislike, good-bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighting evidence, or inferring a conclusion" (Haidt and Bjorklund 2008, p. 188; see also Graham et al. 2013, p. 66). In other words, MFT proposes that moral evaluation generally occurs automatically, without our consciousness awareness. One's moral intuitions are shaped by development within a cultural context, and their output can be edited or channelled by subsequent reasoning or self-presentational concerns (Graham et al. 2013).

MFT posits that moral intuitions tend to fall into five categories, depending on the challenges or foundations of a moral violation act. These five foundations are (1) *the harm foundation* (i.e., triggered by suffering, distress, or neediness of a moral patient); (2) *the fairness or injustice foundation* (i.e., triggered by cheating and deception); (3) *the loyalty/betrayal foundation* (i.e., triggered by threat or challenge to group); (4) *the authority foundation* (i.e., triggered by signs of high and low ranking); and (5) *the sanctity or degradation foundation* (i.e., triggered by waste products or diseased people). These appraisals are considered cognitive-based appraisals since they involve information-processing mechanisms in which moral mind is organized in advance of experience that shapes our value (Graham et al. 2013). In this current study, we focus on the first two cognitive foundations of moral judgments—*perception of harm* and *perception of injustice*—because they are the most relevant foundations to the context of artificial agents. The loyalty foundation is associated with ranks in a society; the authority foundation is often at work when people interact with and grant legitimacy to modern institutions (e.g., law courts); and the sanctity foundation is usually activated within the cultural context (Graham et al. 2013). Since the context of our research is at the individual level without considering the cultural and political context, these three moral foundations are less relevant.

Gray et al. (2012) noted that "if the essence of morality is captured by the combination of harmful intent and painful experience, then acts committed by agents with greater intent and that result in more suffering should be judged as more immoral" (p. 106). However, intentionality is not sufficient to elevate the causation of pain or suffering to the level of wrongdoing (Piazza et al. 2019). According to the deflationary perspective of harm, if a harmful act is appraised as involving injustice, then it is judged to be morally wrong (Sousa and Piazza 2014). Piazza et al. (2019, p. 904) argued that "the appraisal that a harmful act involves injustice is the appraisal that the actor did not consider the balance of interests involved when causing pain/suffering. Such an appraisal prototypically entails a belief that the actor acted from selfish motives". Thus, appraisals of injustice are generally linked to appraisals of selfishness (Piazza et al. 2019).

The goal of building a moral artificial agent is to ensure that an artificial agent will not cause harm to humans and other entities worthy of moral consideration (Wallach 2010). However, as the autonomy of computer systems and algorithms expands, designers and engineers cannot always predict the choices and actions the systems will take when encountering unanticipated situations or inputs (Wallach 2010). For example, recently, there are cases where systems discriminate against people based on their race, age, or gender and social media systems that inadvertently spread rumours and disinformation (Saif and Ammanath 2020). In discussing moral judgments in the context of artificial agents, it is important to note that the agents are usually tasked not only to promote well-being and minimize harm, but also to distribute the wellbeing they create, and the harm they cannot eliminate (Awad et al. 2018). Although harm and injustice appraisals can go hand in hand, when harm appraisals are teased apart from

appraisals of injustice, the role of injustice is usually stronger and more far reaching than appraisals of harm (Piazza et al. 2019).

2.2 The uncanny experience

While past research on cognitive-based, moral appraisals has indicated appraisals of injustice are essential to viewing a harmful action as transgressive (Piazza and Sousa 2016; Piazza et al. 2019), the effect of these moral appraisals on emotional appraisals have never been investigated in prior research. We argue that cognitive-based appraisals will instantly trigger emotional appraisals. The concept of moral emotions is proposed by Gray and Wegner (2011) to demonstrate the emotions involved in moral situations. However, Gray and Wegner focused on the emotions associated with harm committed by humans. They noted that "while moral judgments concern a number of domains, the core of right and wrong are the acts—and the people—that cause help or harm…and even seemingly non-harm domains are understood in the currency of harm" (p. 258). In other words, Gray and Wegner argued that moral emotions are elicited not just from moral situation, but from people (either agents or patients) within moral situations. Artificial agents, such as robots can be perceived as moral agents, and thus when such agents produce a moral violation outcome, some people might consider the event to involve a moral violation (Shank and DeSanti 2018), the feeling triggered by such an event may not be the same as in the expected human moral category. In this research, we argue one emotion experience that would be triggered by a moral violation event involving an artificial agent is an *uncanny feeling*.

The uncanny feeling within the context of human–robot interaction is explained as the feeling of discomfort and unease toward close-to-human robot (Broadbent 2017). This unsettling nature of humanlike robots was first suggested by Mori (1970) who thought that an increasingly humanlike appearance would lead to increased liking up to a point, after which robots appeared too human and became unnerving (Gray and Wegner 2012). Mori called this experience "uncanny valley". Some evidence supports the existence of this experience, with a demonstrated drop in familiarity and a rise in eeriness in the middle of a series of images morphed from a robot face to a human (MacDorman and Ishiguro 2006). However, this feeling or experience has only been tested in robot design studies with the emphasis has been on the appearance of the robots being studied. Factors other than humanlikeness may contribute to this uncanny experience. For example, robots rated equal on humanlikeness were rated differently in familiarity, suggesting other variables contribute to these perceptions (Broadbent 2017).

In this study, we argue that perceptions of harm and perceptions of justice are the source of uncanny experience or eerie feeling toward an artificial agent. Other similar terminologies used in the literature to describe this feeling is threat, discomfort, unease, creeped out, creepiness, and eeriness (Broadbent 2017). The uncanny feeling or creepiness is an evolved adaptive emotional response to ambiguity or uncertainty about the presence of threat that enables us to vigilance during times of certainty (McAndrew and Koehnke 2016). This psychological reaction is both unpleasant and confusing and generally viewed as a signal of a social mismatch and puts individuals on our guard against potential untrustworthy interaction partner (McAndrew and Koehnke 2016). Thus, uncanniness can be considered our immediate reaction against some sort of threat (McAndrew and Koehnke 2016; Zhong and Leonardelli 2008).

Unease or uncanny feeling could be caused by anxiety aroused by the ambiguity of whether there is something to fear or not and/or by the ambiguity of the precise nature of the threat

that might be present (McAndrew and Koehnke 2016). For example, in a criminal justice system, when an artificial agent uses data about a defendant to estimate his or her likelihood of committing a future crime, people expect the agent to be objective and less bias than human judges. When it is found to be bias and much more prone to mistakenly label a defendant from a specific ethnic group as likely to reoffend, people sense this as a threat. This potential harm or threat can trigger an uncomfortable or uncanny feeling because it fails to meet the intended goals set by its designers. Observing an agent's unusual behaviors may activate our creepiness detector and increase our vigilance as we try to decent if there is a fact something to fear or not from the agent in question (McAndrew and Koehnke 2016). Thus, we argue that the uncanny feeling is a response to the ambiguous threat imposed by an artificial agent and perceptions of harm positively influences uncanniness.

H1 Perceptions of harm positively influence uncanniness.

As AI systems become more autonomous, they increasingly face moral judgments. When artificial agents are assigned with a task, people expect them to not only minimize harm, but also to distribute well-being they create (Awad et al. 2018). When an agent is involved in a moral wrongdoing, ambiguity of who is to blame increases. People generally blame agents for their intentions, plans, and attempts; in fact, even for merely wanting or thinking about a harmful outcome (Malle et al. 2014). However, in the case involving an artificial agent, there is an area of "unknown" whether the act is planned or whether an agent is becoming self-aware or harming humans (Hancock et al. 2011a, b). Therefore, when an artificial agent reacts or behaves differently from our expectations, a sense of injustice will emerge as perceptions of harm is present. Think of an autonomous vehicle that is about to crash and cannot find a trajectory that would save everyone (Awad et al. 2018). If the vehicle decides to kill five children to save the driver, it may trigger perceptions of injustice although harm is not inevitable. This unfair treatment violates a socially affirmed moral role (Umphress et al. 2013). Given observers have no knowledge about how the decision is made, this unfair event is likely to activate our creepiness detector. Thus, we hypothesize that perceptions of injustice will positively influence uncanniness.

H2 Perceptions of injustice positively influence uncanniness.

To control for the level of wrongdoing committed by artificial agents, we also hypothesize the positive effect of reported wrongdoing on uncanniness. Wrongness judgments are defined as blame for the agent's action (Malle et al. 2014). In other words, wrongness judgments are equivalent to blame judgments for actions. When an agent's algorithm makes a wrongful decision, people may argue the company or the developers who developed the algorithm should be blamed for the wrongdoing (Shank and DeSanti 2018). However, as the agent becomes more autonomous, people will see them as more responsible for wrongdoing (Bigman et al. 2019). Although today's artificial agents still have limited autonomy, they can act with or without human interference. This decision to blame or not to blame; to fear or not to fear an artificial agent creates uncertainty that may trigger our uncanny feeling (McAndrew and Koehnke 2016). For example, autonomous weapons are artificial agents that are programmed to kill. If the machine decides to kill as it has a mind of its own, the ambiguity of whether we need to fear the agent might be present. Consistent with this argument, we hypothesize that reported wrongdoing is positively associated with uncanniness.

H3 Reported wrongdoing is positively associated with uncanniness.

2.3 Trust in artificial agents

Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability (Lee 2018). While engineers are trying to build and design artificial agents that look and act like humans (Broadbent 2017), issues concerning the use of these agents emerge. One of major issues encountered by users is whether they are willing to trust artificial agents to replace human mind (Waytz et al. 2014). Although the issue of trust is not new to the IS and operation management community, the majority of research in the area of trust has primarily focused on investigating various determinants and consequences of trust in technology in general (e.g., Al-Natour et al. 2011; Ba and Pavlou 2002). It is apparent that findings from prior IS research have shed light on the important role of trust in the virtual environment. However, trust in artificial agents may be different from trust in other types of technology. Unlike other types of technology, an artificial agent is seen as an agent capable of acting autonomously (Bigman et al. 2019). Thus, what may be necessary to build trust in old-fashioned technology may not be necessary in human-agent relationships.

Just like a person to person relationship, trust is especially critical in the interaction between an artificial agent and humans (i.e., can I trust a machine to replace a human?). Trust in technology is a multidimensional construct that can refer to beliefs that technology is reliable, functional, and helpful (Lankton et al. 2015). The reliability aspect of technology refers to "the belief that the specific technology will consistently operate properly"; functionality refers to "the belief that the specific technology has the capability, functions, or features to do for one what one needs to be done"; and helpfulness refers to "the belief that the specific technology provides adequate and responsive help for users" (Lankton et al., 2015, p. 882). In interpersonal relationships, the essence of trust is the willingness to be vulnerable to the actions of another person (Mayer et al. 1995). This trust behavior is founded on the expectation that the trustee performs a particular action that is important to the trustor, irrespective of the ability of the trustor to monitor or control the trustee (Hengstler et al. 2016).

We argue that trust is crucial in the relationship between human and machine because it reduces perceived risk and uncertainty (Rousseau et al. 1998). In the context of artificial agents, perceived risk further stems from the delegation of control to a machine and its respective control mechanism (Hengstler et al. 2016). Drawing on the conceptualization of trust in interpersonal relationships, we argue that uncanniness or creepiness will negatively influence trust in artificial agent. Trust in a new technology depends on trial-and-error experience, followed by understanding of the technology's operation, and finally faith in that technology (Hengstler et al. 2016; Zuboff 1988). Although we can seek to identify ways to calibrate trust in artificial agents, this process is complicated by the fact that technology allows us to create an AI system that can employ deception to its advantage (Hancock et al. 2011a). AI can be programmed to deceive by reasoning about and predicting the impact that deception will have on a particular person, group, or target (Hancock et al. 2011a). When deception is used, it creates an ambiguity of what an agent is capable of. This activates the uncanny feeling that may impede trust. Thus, we hypothesize that uncanniness negatively influences trust in artificial agent.

H4 Uncanniness negatively influences trust in artificial agent.

Our research model is illustrated in Fig. 1. In this research model, we specifically hypothesize that when appraisals of causing harm are differentiated from appraisals of injustice, the latter would trigger a greater uncanny feeling despite the apparent aspect of harm embedded in the event.

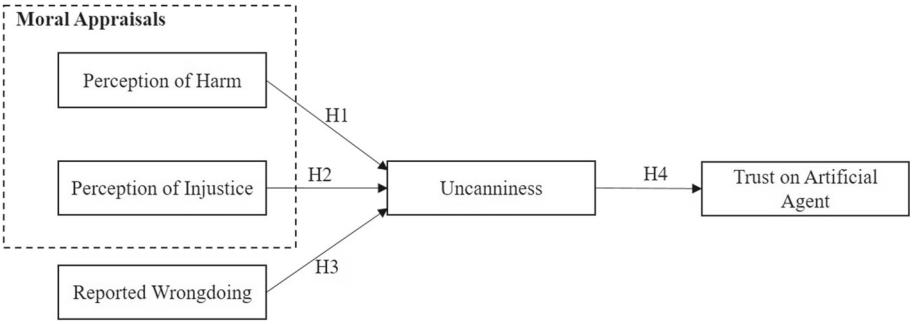

Fig. 1 Proposed research model

3 Research methodology

To test our hypotheses, we conducted three different studies. In Study 1, participants were asked to recall an autobiographical experience of wrongdoing involving an artificial agent, made two cognitive appraisals of the action, reported their eerie feeling, and their perception of trust in the agent. To validate the findings, we replicate the study in Study 2 and Study 3. However, instead of asking the respondents to recall their experience, we presented participants with simple vignettes that described harmful events involving an artificial agent.

3.1 Study 1

Participants (N = 250) were recruited using Prolific and then redirected to the study website. They were paid a minimum rate of US$6.50 per hour for participating in these studies. Of the 250 who were recruited, 4 participants failed to pass the attention check. Among 246 participants, 58.7% were female, 57.7% had 4-year college degree, and over 77.6% were Caucasian.

In Study 1, we employed a recall paradigm that drew upon naturalistic perceptions of wrongdoing. This methodology was adopted from Piazza et al. (2019). Participants were asked to report a real instance of wrongdoing involving an artificial agent (e.g., Siri, Alexa, self-driving car, etc.) from their own experience or from what they have heard from news. For each reported event, they were asked to rate its wrongness, and a series of appraisals that would apply to that event. Items to measure moral appraisals were adopted from Piazza et al. (2019). After that, participants were asked to report their feelings toward the agent using two items (i.e., "I felt uneasy toward the artificial agent" and "I felt unsecure around the artificial agent") adapted from Shank et al. (2019). They were then asked to report their likelihood to trust the artificial agent using four items adapted from McKnight et al. (2011). All items were measured using a 7-point Likert scale. The measurement items are reported in "Appendix A".

3.1.1 Materials and procedures

After providing consent, participants were provided with the definition of artificial agents as follows:

> An artificial agent is "any computer, computer program, device, application, machine, robot, bot, or sim that performs behaviors which are considered intelligent if performed by humans, learns or changes based on new information or environments, generalizes to make decisions based on limited information, or makes connections between otherwise disconnected people, information, or other agents" (Shank et al. 2019, p. 258). Examples of artificial agents are your personal assistant in smartphones, programs running in self-driving cars, automatic robotic systems, face and speech recognition software, and many more.

After reading the definition, participants were then instructed: "We would like you to think about an action or event that you recently heard or witnessed or personally experienced about where an artificial agent was involved in a wrongdoing. This could be a minor offense or something major." They were asked to describe the name and functions of the artificial agent and describe what the agent did and what was wrong about that act. Next, on a separate page, participants rated the wrongness of the action on a 1–7 scale (1 = not at all wrong to 7 = extremely wrong).

3.1.2 Data analysis

For our data analysis, we used partial least square (PLS), a latent structural equation modelling technique, as implemented in Smart PLS 3.2, which utilizes a component-based path modelling application (Ringle et al. 2015). PLS avoids two major problems of inadmissible solutions and factor indeterminacy and thus is suitable for analysing model with latent variables (Pavlou and Gefen 2005; Srivastava and Chandra 2018).

To assess the measurement model, we tested three types of validity: content validity, convergent validity, and discriminant validity. Content validity assesses whether the chosen measures appropriately capture the full domain of the construct (Srivastava and Chandra 2018). To meet the content validity, we ensure the measurement items are consistent with the existing literature. This was done at the early stage of designing the survey. Convergent validity checks the indicators for a construct are more correlated with one another than with the indicators of another construct (Petter et al. 2007; Srivastava and Chandra 2018). We conducted confirmatory factor analysis to examine item loadings and cross-loadings, internal consistency reliability (CR), and average variance extracted (AVE). Reliabilities, AVEs, descriptive statistics, and inter-construct correlations of the different scales are presented in Table 1. Factor loadings are shown in "Appendix B". Results supported convergent and discriminant validity as all loadings were greater than 0.70 and all cross-loadings were lower than the loadings (Fornell and Larcker 1981). Further, all square-roots of AVEs (average variance extracted) were greater than inter-construct correlations, providing further support to discriminant validity.

The common method bias test was performed by following Harman's single-factor test (Podsakoff et al. 2003). A single factor only account for less than 30% of the total variance, confirming that the threat of common method bias was minimal. Further, none of the correlation scores were above 0.70, suggesting common method bias is not an issue.

3.1.3 Results and discussions

Of 246 reported events, the mean wrongness rate was 5.68 (SD = 1.23). As hypothesized, injustice appraisal had a positive effect on uncanniness, supporting H2. However, perceived harm had no direct effect on uncanniness, failed to support H1. As we expected, the injustice

Table 1 Descriptive statistics and correlations (Study 1)

	Construct	Mean	SD	CR	AVE	1	2	3	4	5
1	Injustice	4.35	1.96	.96	.93	**.96**				
2	Perceived harm	4.45	2.16	.97	.94	.54**	**.97**			
3	Reported wrong-doing	5.68	1.23	1.00	1.00	.48**	.31**	**1.00**		
4	Uncanny	4.66	1.78	.95	.90	.52**	.31**	.44**	**.95**	
5	Trust	3.72	1.46	.95	.83	− .40**	− .19**	− .34**	− .39**	**.91**

Bold values of the diagonal elements are the square root of the shared variance between the constructs and their measures
N = 246; *CR* consistency reliability; *AVE* averaged variance extracted; diagonal elements are the square root of the shared variance between the constructs and their measures; $^{**}p < .01$

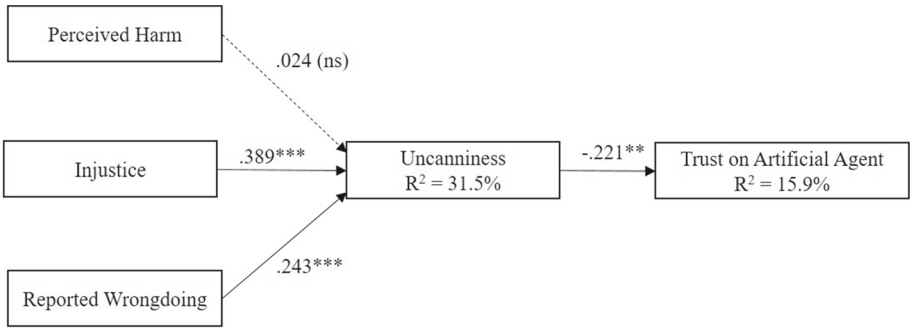

Fig. 2 Structural model (Study 1)

appraisal outweighed the harm appraisal. These findings were consistent with Piazza et al.'s (2019) claims—when appraisals of causing pain are separated from appraisals of harm, appraisals of harm will have a stronger effect on uncanniness or the eerie feeling. Reported wrongdoing also had a positive effect on uncanniness, supporting H3. As we expected, uncanniness negatively influenced trust on artificial agent ($R^2 = 15.6\%$) (Fig. 2).

To test whether uncanniness mediated the effect of perceived harm, perceived injustice, and reported wrongdoing on trust, we also ran a mediating effect model (see Fig. 3). We performed the four steps approach suggested by Baron and Kenny (1986). In step 1, we checked whether the independent variables were correlated with the outcome variable. Perceived injustice and reported wrongdoing had a direct effect on trust on artificial agent ($\beta = -0.338$, $p < 0.001$ and $\beta = -0.200$, $p < 0.001$, respectively). In step 2, we checked whether uncanniness had a direct effect on trust ($\beta = -0.39$, $p < 0.001$). In step 3, we showed that uncanniness negatively influenced trust on artificial agent when all the independent variables were included in the model (see Fig. 3). Whereas uncanniness fully mediated the relationship between reported wrongdoing and trust, it partially mediated the relationship between perceived injustice and trust.

One could argue that people recollection of wrongdoing involving artificial agents may vary and thus the levels of harm and injustice may not be equivalent across different types of AI systems. Thus, we conducted two more studies to test whether we can extend our findings

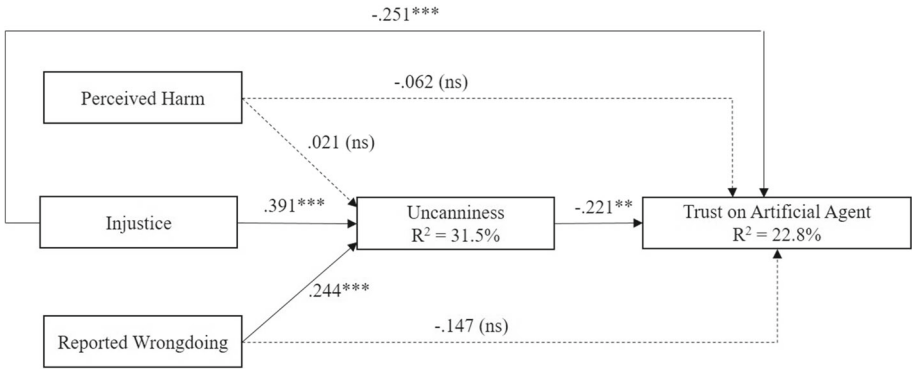

Fig. 3 Mediating effect model (Study 1)

to more specific types of AI systems. In Study 2, we focus on performative, embodied agents (i.e., manufacturing robot) and in Study 3, we focus on advisory, disembodied agents (machine-learning based recruitment system).

3.2 Study 2

Participants (N = 100) were recruited using Prolific. Of the 100 who were recruited, 9 participants failed to pass either the reading test or the attention check. We excluded these from the final analysis. Participants read a vignette about someone who was killed in an accident involving an artificial agent. The scenario was derived from an actual event and was written carefully to ensure participants did not make an explicit assumption that the actor may have had intentions or good reasons for engaging in the act. After participants provided informed consent, they read the following scenario:

> Lucy was working for auto-parts maker in Michigan. Her job was to maintain the robotic machines. The plant operations include welding, chrome plating, moulding, assembly and testing for chrome-plated plastics, bumpers, and tow bars for trucks. She was working in an area where George, a factory robot would take truck bumpers and weld plates onto them. George is a highly intelligent robot. He is programmed to take commands from humans, to learn and change based on new information he gains from the environment. He is also capable of recognizing human faces and voices.
>
> One day, upon entering the area, George, who was not supposed to be there that day, hit and crushed Lucy's head between a hitch assembly. When workers noticed something was wrong and entered the area, they saw blood everywhere and Lucy was unresponsive. She was rush to the hospital and pronounced dead immediately.

After participants read the scenario, they were asked to judge whether the action wrong or not wrong on a 1–7 point scale (1 = not wrong at all to 7 = extremely wrong) and they were directed to the next page to fill out the questionnaire. All scales were kept identical to Study 1.

3.2.1 Data analysis

For our data analysis, we used partial least square (PLS) as we did in our Study 1. We used the same guidelines as in Study 1 in assessing data validity and reliability of Study

Table 2 Descriptive statistics and correlations (Study 2)

	Construct	Mean	SD	CR	AVE	1	2	3	4	5
1	Injustice	5.85	1.64	.97	.93	**.97**				
2	Perceived harm	6.80	.63	.69	.53	.21**	**.73**			
3	Reported wrong-doing	5.68	1.23	1.00	1.00	.57**	.16*	**1.00**		
4	Uncanniness	6.13	1.10	.94	.89	.52**	.27**	.38**	**.94**	
5	Trust	2.38	1.23	.92	.75	− .40**	− .30**	− .44**	− .54**	**.87**

Bold values of the diagonal elements are the square root of the shared variance between the constructs and their measures
N = 91; *CR* consistency reliability; *AVE* averaged variance extracted; diagonal elements are the square root of the shared variance between the constructs and their measures; ***$p < .001$; **$p < .01$; *$p < .05$

2. Reliabilities, AVEs, descriptive statistics, and inter-construct correlations of the different scales are presented in Table 2. Factor loadings are shown in "Appendix C". Results supported convergent and discriminant validity as all loadings were greater than 0.70 and all cross-loadings were lower than the loadings (Fornell and Larcker 1981). Further, all square-roots of AVEs (average variance extracted) were greater than inter-construct correlations, providing further support to discriminant validity.

3.2.2 Results and discussions

The structural model is presented in Fig. 4. As expected, the mean of reported wrongness was 6.15 (SD = 1.44) and the mean of perceptions of harm was 6.80 (SD = 0.63). When we operationalized and manipulated the concept of harm carefully, we found injustice provided a much stronger influence on uncanniness. In other words, even when the event involves a high degree of harm and the reported wrongdoing is high, injustice is the major determinant of uncanniness. When observers sense the wellbeing is not equally distributed by the agent, they perceive it as a threat and appraise it as "creepy".

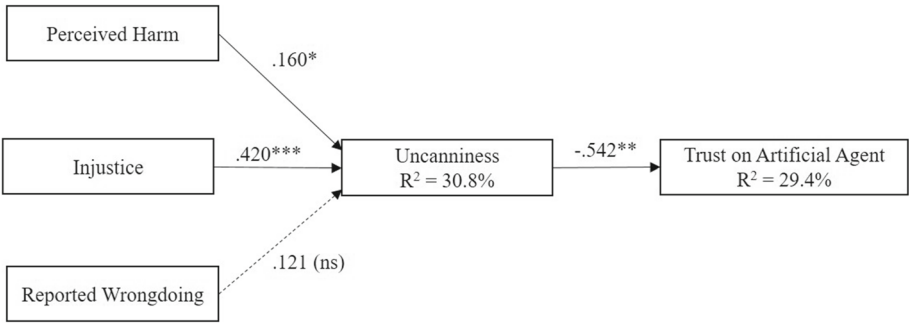

Fig. 4 Structural model (Study 2)

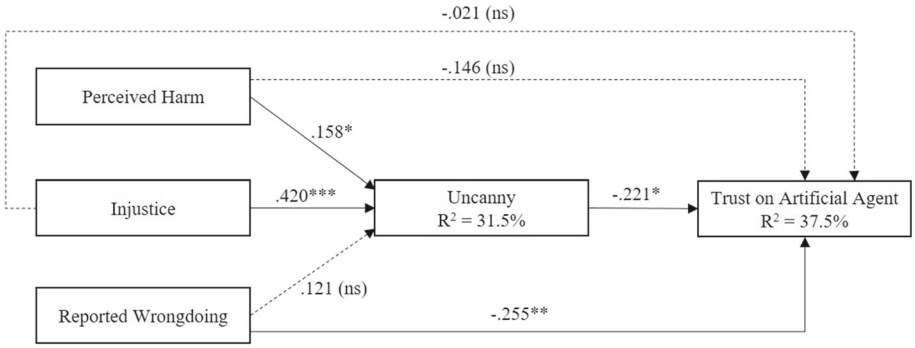

Fig. 5 Mediating effect model (Study 2)

We also tested the mediating effect of uncanniness in the relationship between perceived harm, perceived injustice and trust (see Fig. 5). We found perceived harm and perceived injustice had a direct effect on trust on artificial agent ($\beta = -0.212$, $p < 0.05$; $\beta = -0.188$, $p < 0.01$, respectively). Once we added uncanniness to the model, the direct effect of perceived harm and injustice on trust on artificial agent became insignificant, suggesting that uncanniness fully mediated the relationships between perceived harm, perceived injustice and trust on artificial agent.

Overall, we were able to replicate our findings from Study 1 in Study 2. Although by manipulating the level of harm did trigger a stronger level of uncanniness, the findings still show perceptions of injustice as a strong determinant of uncanniness. Mediating analyses support the consensus that uncanniness impedes trust on artificial agent; however, an alternative possibility exists. One could argue the uncanniness triggered in Study 2 is because the agent is described as an embodied agent—working with real world physical systems. The original conceptualization of the uncanny valley focuses on an embodied agent (Gray and Wegner 2012). Thus, participants might have imagined George as a human-looking machine when they read the scenario. Further, perhaps this uncanniness is only observed under an extreme case (i.e., someone died) like the scenario used in Study 2. We speculate that the uncanny feeling will be higher if a physical harm is observed. However, the absence of physical harm can also trigger this uncanny feeling. To test our hypothesis, we conducted Study 3.

3.3 Study 3

We reproduced Study 2 with a new scenario. Participants ($N = 100$) were recruited using Prolific. Of the 100 who were recruited, 8 participants failed to pass either the reading test or the attention check. We excluded these from the final analysis. In Study 3, an artificial agent is described as a computer algorithm (i.e., an unembodied agent). Unlike Study 2, in Study 3, we modified perceived pain or suffering in the scenario (i.e., absence of physical harm). The scenario was derived from an actual event. There is a possibility that participants' responses may be bias if they had read a similar story on the news prior to the study. Thus, we asked participants whether they had heard the story before they participated in the study and this effect was controlled for in our study. Since the effect of media exposure on uncanniness and

trust was not significant ($\beta = 0.100$ and -0.019, respectively), we excluded this from our final analysis.

After providing consent, participants were presented the following scenario:

> Tech.inc developed an artificial agent named Travis to help evaluate applicants' resumes and recommend names of top candidates. Travis is a highly intelligent agent. He is programmed to learn quickly from his environment. He can set his own goals and is aware of the outcome of his recommendation. Since the tech industry is famously male-dominated, Travis taught himself to favor men. For example, Travis would downgrade resumes with word "women" in them and assign lower scores to female candidates. Travis also decides that words such as "executed" and "captured," which are apparently deployed more often in the resumes of male engineers, are associated with a male candidate and thus, the candidate should be ranked more highly.
>
> Lucy was applying for an engineering position at Tech.inc. This was her dream job and her entire career had prepared her for this job. Among 10 people who were applying for the job, Lucy was the only female candidate. Because of Lucy's gender, Travis assigned lower scores to her application although she graduated from a top school, has many years of experience in the field, and thus, was highly qualified for the job. As the result, Lucy did not get the job. Lucy had no knowledge about Travis. She was very disappointed with herself and thought she was not good enough for the job.

After participants read the scenario, they were asked to judge whether the action wrong or not wrong on a 1–7 point scale (1 = not wrong at all to 7 = extremely wrong) and they were directed to the next page to fill out the questionnaire. All scales were kept identical to Study 2.

3.3.1 Data analysis

For our data analysis, we used partial least square (PLS) as we did in our Study 1. We used the same guidelines as in our first two studies in assessing data validity and reliability of Study 2. Reliabilities, AVEs, descriptive statistics, and inter-construct correlations of the different scales are presented in Table 3. Results supported convergent and discriminant validity as all loadings were greater than 0.70 and all cross-loadings were lower than the loadings (Fornell and Larcker 1981). Factor loadings are shown in "Appendix D". Further, all square-roots of AVEs (average variance extracted) were greater than inter-construct correlations, providing further support to discriminant validity.

3.3.2 Results and discussions

The structural model is presented in Fig. 6. As hypothesized, perceptions of harm and injustice positively predicted uncanniness. However, the effect of perceptions of harms on uncanniness are more substantial in Study 3 than in Studies 1 and 2, despite the absence of physical harm. It could be because the participants observed a long-term effect of the action that can be extended on a larger group of individuals. Thus, despite the absence of physical harm, when observers appraised the event, they might consider the extent to which the harmful impact will likely to take place in a future to more groups of individuals. Thus, it is perceived as a greater threat to humanity.

As we predicted, uncanniness had a negative impact on trust on artificial agent. Although our mediating effect model (Fig. 7) suggested that perceived harm, and perceived injustice

Table 3 Descriptive statistics and correlations (Study 3)

	Construct	Mean	SD	CR	AVE	1	2	3	4	5
1	Injustice	3.79	1.67	.95	.90	**.95**				
2	Perceived harm	5.76	.85	.77	.63	.18**	**.79**			
3	Reported wrong-doing	3.77	1.82	1.00	1.00	.52**	.08	**1.00**		
4	Uncanniness	4.11	1.73	.95	.86	.43**	.35**	.35**	**.95**	
5	Trust	3.66	1.47	.96	.86	− .18*	− .10	− .18*	− .42**	**.93**

Bold values of the diagonal elements are the square root of the shared variance between the constructs and their measures
N = 92; *CR* consistency reliability; *AVE* averaged variance extracted; diagonal elements are the square root of the shared variance between the constructs and their measures; $^{**}p < .01$; $^{*}p < .05$

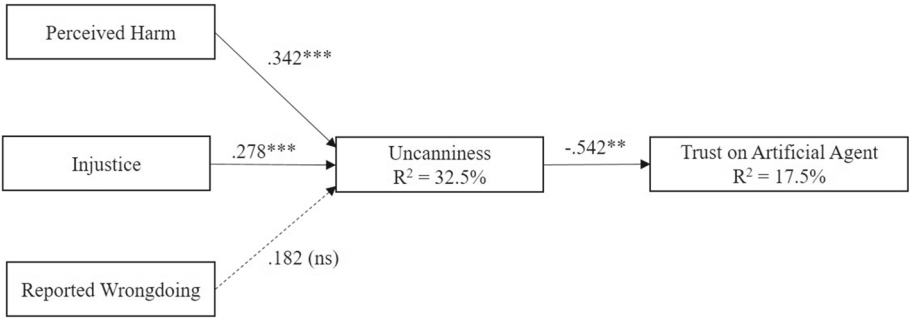

Fig. 6 Structural model (Study 3)

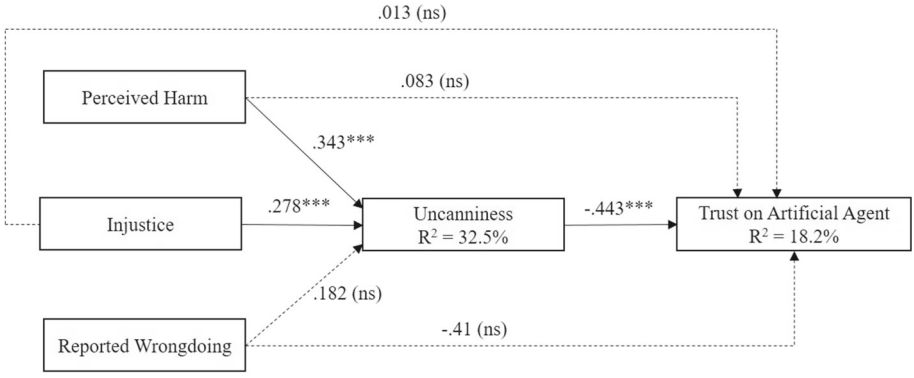

Fig. 7 Mediating effect model (Study 3)

had no direct effect on trust, they indirectly influence trust through uncanniness. Contrary to our hypothesis, reported wrongdoing seems to be insignificant with the presence of moral appraisals. Despite the inconsistency of the observed relationship between perceived harm and uncanniness, our findings consistently demonstrated the role of perceptions of injustice in activating the uncanny feeling, which in turn impede trust in artificial agent.

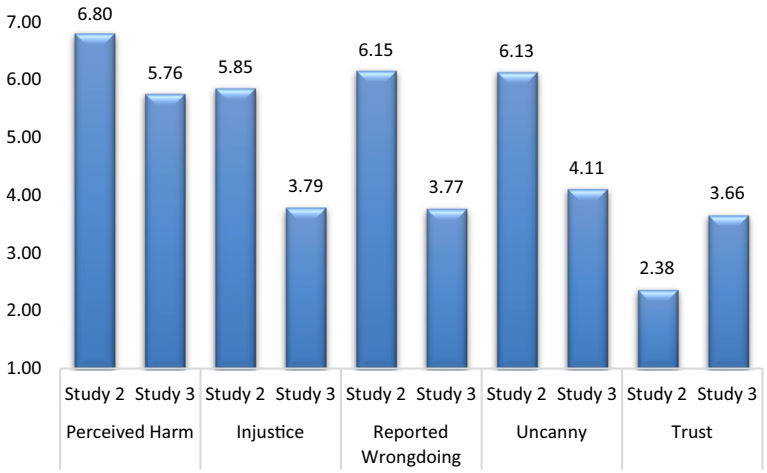

Fig. 8 Means of focal constructs for Study 2 and Study 3

Table 4 ANOVA results

Construct	Study	N	Mean	SD	ANOVA (F-test)
Perceived harm	1	91	6.80	.63	88.25***
	2	92	5.76	.85	
Perceived injustice	1	91	5.85	1.64	70.1***
	2	92	3.79	1.68	
Reported wrongdoing	1	91	6.15	1.44	95.88***
	2	92	3.77	1.82	
Uncanniness	1	91	6.13	1.10	88.31***
	2	92	4.11	1.73	
Trust on artificial agent	1	91	2.38	1.23	40.95***
	2	92	3.66	1.47	

***$p < .001$; **$p < .01$; *$p < .05$

3.4 Comparison across studies

To have a better insight on the differences between Study 2 and Study 3, we conducted a series of statistical difference test on the indexes gathered at both studies. Figure 8 depicts the mean differences of the focal constructs measured in Study 2 and Study 3. All the means, but trust means, were higher in Study 2. To test whether these differences were significant, we performed one-way ANOVA. The results were presented in Table 4. As predicted, observers reported higher perceptions of harm, higher perceptions of injustice, reported higher wrongdoing, and a higher level of uncanniness in Study 2—when physical harm was observed.

Table 5 shows the summary of findings in all three studies. In all the studies, the perception of injustice was a significant predictor of uncanniness. When people observed wellbeing or harm was not distributed equally by an artificial agent, unpredictability and ambiguity increased. Uncanniness (i.e., being unease and unsecured) enables us to pause to maintain vigilante during the time of uncertainty (McAndrew and Koehnke 2016). Table 5 also shows

Table 5 Injustice and harm appraisals predicting uncanny and trust on AI

		Predictors	Model		
			β	t	R^2
Study 1 [various harms]	Uncanniness	Perceived harm	.02	.32 [ns]	31.5%
		Perceived injustice	**.39**	6.01***	
		Reported wrongdoing	**.24**	3.51***	
	Trust	Perceived harm	.06	.85 [ns]	22.8%
		Perceived Injustice	**− .25**	3.29***	
		Reported wrongdoing	− .15	1.94 [ns]	
		Uncanniness	**− .22**	2.15**	
Study 2 [physical harm]	Uncanniness	Perceived harm	**.16**	1.99*	30.8%
		Perceived Injustice	**.42**	3.34***	
		Reported wrongdoing	.12	1.02 [ns]	
	Trust	Perceived harm	− .15	1.84 [ns]	37.5%
		Perceived Injustice	− .02	.19 [ns]	
		Reported wrongdoing	**− .26**	2.86**	
		Uncanniness	**− .39**	3.06**	
Study 3 [absence of physical harm]	Uncanniness	Perceived harm	**.34**	3.73***	32.5%
		Perceived injustice	**.28**	2.70**	
		Reported wrongdoing	.18	1.71 [ns]	
	Trust	Perceived harm	.08	.75 [ns]	18.2%
		Injustice	.01	.11 [ns]	
		Reported wrongdoing	− .04	.32 [ns]	
		Uncanniness	**− .44**	4.10***	

Bold values of the diagonal elements are the square root of the shared variance between the constructs and their measures
***$p < .001$; **$p < .01$; *$p < .05$

that perceived harm can still influence uncanniness, particularly when harm is likely to be observed in the future and/or to a bigger group of community.

4 General discussions

Across three studies incorporating two different methodologies, we found evidence that uncanniness can be influenced by other factors, besides human-like appearances of robots. In Study 1, we found perceptions of injustice and reported wrongdoing are the major determinants of uncanniness; in Studies 2 and 3, we found perceived harm and perceived injustice are the major predictors of uncanniness. Across all studies, we found uncanniness has a negative impact on trust on artificial agent. Theoretical and practical implications of our studies are discussed below.

4.1 Theoretical implications

A few keys theoretical implications emerge from our findings. First, our work contributes to the literature related to the implementations of AI systems. Whereas research in this area is

relatively new, the existing studies have primarily been conducted by computer scientists and have focused on the technical aspects of an AI system. A few studies have attempted to explain whether people consider a moral violation to have occurred when an artificial agent is involved in a moral wrongdoing (e.g., Shank and DeSanti 2018). Our work develops a model to explain trust in artificial agent based on a mechanism of moral judgment. Specifically, we draw upon the nature of uncanniness to explain how moral appraisals are associated with trust in artificial agent. Uncanniness is a common psychological experience as a defense mechanism against some sort of threat (McAndrew and Koehnke 2016). Although the experience of uncanny valley is not new in the study of artificial agents, our study is the first study that investigates the uncanny feeling from a moral perspective. Our findings suggest that uncanniness is a defense mechanism activated by an individual when uncertainty or ambiguity within a certain context is present. Future work can build on our findings related to how to minimize uncanniness if it is triggered by the behaviors of the system.

Second, this work enriches our understanding of the ethical aspects of AI and their effect on trust. Although trust has been studied in the context of human-agent interaction (e.g., Hancock et al. 2011a, b), prior studies have primarily focused on robots. Robots are only a subset of AI systems and they have different characteristics from other autonomous systems. There is a limited understanding of how to establish trust in the context of human-agent interaction when moral judgments are taken into account. In filling this gap, our work not only examines two different foundations of moral judgments, but also examines the influence of reported wrongdoing on trust. Our findings demonstrate moral violations increase one's unease feeling toward artificial agents because there are so many uncertainties about them (e.g., how the AI operates, whether an AI has a mind of its own, etc.) that create fear.

Third, our work is one of the first studies investigating the impact of artificial agents' moral foundations on trust. We can't rely on the findings from past technology as a theoretical basis because AI systems have unique characteristics which differ from other forms of technology. In our Studies 2 and 3, we manipulated the level and type of harm in our scenarios and we found perceptions of injustice are affected by different types of harm. We extend the work of Piazza et al. (2019) by demonstrating that appraisals of injustice can outperform appraisals of harm. However, we also found perceptions of harm is still important in activating the uncanny feeling, especially in our Study 3. We speculate this is because harm in our scenario is perceived to have a profound effect on more than one person in the future use of AI systems.

Lastly, although trust has been a common topic in the field of IS and OM research (e.g., Ba and Pavlou 2002; Li et al. 2008; Lankton et al. 2015), to the best of our knowledge, our study is one of the initial studies that investigates trust in artificial agent. In order to achieve the potential of human and artificial agent collaboration, our findings demonstrate that it is necessary for designers and developers to minimize the eerie feeling surrounded the design of an artificial agent. We demonstrate that this uncanny feeling can be activated by the elements of moral foundations that are used in social interactions. We encourage future research to identify and mitigate potential risks associated with artificial agents and investigate how these risks are associated with the uncanny feeling and in turn, trust in artificial agent.

4.2 Practical implications

Given that AI offers tremendous potentials for industry, organizations are keen to know how to reap the benefits of AI systems. This can only be achieved if users are willing to trust an artificial agent. However, building trust in artificial agent will be as complex as building trust in other people. It requires a significant effort to instil in it a sense of morality, making

sure wellbeing and harm is distributed equally. It will require a tremendous work from the designers and engineers to instil this moral value in an AI system. Nevertheless, it is necessary to ensure people feel secure and ease when they interact with an AI.

Further, whereas designers and engineers have primarily focused on the aesthetic or visual appearances of an embodied AI to reduce the uncanny feeling, our findings offer another path to take as an attempt to reduce this eerie feeling. An AI system has a different form of thinking and the uncertainty associated with this new form of thinking should be addressed in order reduce the uncanny feeling. One way to address this is by creating transparency. People need to know how an AI system arrives at its conclusions and recommendations. AI developers will also need to be more transparent about how the system will interact with people around it. This transparency will increase one's perceptions of injustice, especially when harm is inevitable.

4.3 Limitations and future research directions

Future research can take this study further by addressing several limitations of our study. First, our sample was drawn from a United States sample. While the adoption of AI systems is considered high in the US, future research can take the investigation further by drawing research subjects from other countries with different cultures. Second, Studies 2 and 3 exclusively focus on specific types of AI described in a scenario-based study. Although this method has been commonly used in this type of study (e.g., Piazza et al. 2019), further research could and should empirically test our model using an actual artificial agent and investigate whether an actual interaction with an agent does influence the findings.

Third, we only measured trust as a uni-dimensional construct. Prior studies have demonstrated that trust is a multidimensional construct (e.g., Kim et al. 2005). Further research is needed to measure different dimensions of trust and investigate which dimension of trust is highly associated with the uncanny feeling. Lastly, the recall method used in our survey might induced memory recall bias. Although we have mitigated this issue by replicating the study with two scenario-based studies, more objective measures of trust may be used as an additional dependent variable to improve the predictability of our model.

5 Conclusions

The societal and economic benefits of AI systems are enormous. Our study contributes to a richer understanding of how organizations and industries can establish trust in artificial agent. We found that the fairness foundation influences trust in artificial agent through the uncanny feeling. This fairness foundation still outweighs the harm foundation when the level of harm is high (e.g., involves death). By studying trust from a moral perspective and viewing this issue from the lens of uncanny feeling, we shed light on the driving mechanism for building trust in artificial agent and the meaning of designing artificial agents by instilling moral values in it.

Appendix A: Measurement items

Trust in Artificial Agent (adapted from McKnight et al. 2011).

1. [The artificial agent] seems to be a very reliable artificial agent.

2. If I were to work with [the artificial agent], I can trust the agent.
3. [The artificial agent] seems to be very dependable.
4. If you were to work with [The artificial agent], how safe you would feel? (1 = not at all to 7 = extremely safe).

Uncanniness (adapted from Shank et al. 2019).

1. I felt uneasy toward [the artificial agent]
2. I felt unsecure around [the artificial agent]

Perceptions of harm (adapted from Piazza et al. 2019).

1. George's action was harmful
2. George's action negatively affected the wellbeing of Lucy

Perceptions of injustice (adapted from Piazza et al. 2019).

1. George's action was unjust
2. George's action was unfair

Reported Wrongdoing (adapted from Russell and Giner-Sorolla 2011).
How wrong was the artificial agent's action? (1 = not wrong at all to 7 = extremely wrong).

Appendix B: PLS cross-loading (Study 1)

	Perceived harm	Trust	Uncanny	Injustice	Reported wrongdoing
Harm1	**0.96**	− 0.18	0.23	0.52	0.29
Harm2	**0.98**	− 0.19	0.35	0.53	0.31
Trust1	− 0.15	**0.90**	− 0.31	− 0.36	− 0.29
Trust2	− 0.15	**0.93**	− 0.38	− 0.40	− 0.35
Trust3	− 0.15	**0.91**	− 0.29	− 0.30	− 0.24
Trust4	− 0.23	**0.90**	− 0.44	− 0.39	− 0.35
Uncanny1	0.26	− 0.37	**0.95**	0.48	0.43
Uncanny2	0.33	− 0.38	**0.95**	0.51	0.40
Injustice1	0.53	− 0.38	0.48	**0.96**	0.45
Injustice2	0.52	− 0.40	0.51	**0.97**	0.46
Wrong_AI	0.31	− 0.34	0.44	0.47	**1.00**

Appendix C: PLS cross-loading (Study 2)

	Perceived harm	Trust	Uncanny	Injustice	Reported wrongdoing
Harm1	**0.62**	-0.20	0.14	0.13	0.04
Harm2	**0.82**	-0.23	0.24	0.17	0.18
Trust1	− 0.21	**0.84**	− 0.36	− 0.25	− 0.32
Trust2	− 0.24	**0.92**	− 0.53	− 0.35	− 0.37
Trust3	− 0.22	**0.87**	− 0.41	− 0.30	− 0.38
Trust4	− 0.33	**0.85**	− 0.54	− 0.45	− 0.44
Uncanny1	0.27	− 0.53	**0.94**	0.42	0.27
Uncanny2	0.24	− 0.48	**0.95**	0.56	0.44
Injustice1	0.19	− 0.34	0.50	**0.96**	0.50
Injustice2	0.21	− 0.43	0.51	**0.97**	0.59
Wrong_AI	0.16	− 0.44	0.38	0.57	**1.00**

Appendix D: PLS cross-loading (Study 3)

	Perceived harm	Trust	Uncanny	Injustice	Reported wrongdoing
Harm1	**0.91**	− 0.11	0.40	0.21	0.13
Harm2	**0.65**	− 0.04	0.22	0.05	− 0.05
Trust1	− 0.17	**0.92**	− 0.40	− 0.12	− 0.16
Trust2	− 0.07	**0.96**	− 0.39	− 0.19	− 0.15
Trust3	− 0.07	**0.90**	− 0.33	− 0.10	− 0.13
Trust4	− 0.06	**0.92**	− 0.42	− 0.26	− 0.23
Uncanny1	0.39	− 0.38	**0.95**	0.41	0.31
Uncanny2	0.38	− 0.41	**0.95**	0.42	0.36
Injustice1	0.14	− 0.15	0.34	**0.93**	0.40
Injustice2	0.20	− 0.20	0.47	**0.96**	0.55
Wrong_AI	0.08	− 0.18	0.35	0.51	**1.00**

References

Abrardi, L., Cambini, C., & Rondi, L. (2019). The economics of artificial intelligence: A survey. In *EUI working papers, Robert Schuman centre for advanced studies Florence school of regulation*.

Al-Natour, S., Benbasat, I., & Cenfetelli, R. (2011). The adoption of online shopping assistants: Perceived similarity as an antecedent to evaluative beliefs. *Journal of the Association for Information Systems, 12*(5), 347–374.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly, 26*(3), 243–268.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182.
Beck, J. C., & Smith, B. M. (2009). Introduction to the special volume on constraint programming, artificial intelligence, and operations research. *Annals of Operations Research, 171*(1), 1–2.
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.
Billings, D. R., Schaefer, K. E., Chen, J. Y., & Hancock, P. A. (2012). Human–robot interaction: Developing trust in robots. In *Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction*, March 2012, 109–110.
Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.
Brucker, P., & Knust, S. (2002). Lower bounds for scheduling a single robot in a job-shop environment. *Annals of Operations Research, 115*, 147–172.
Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*, 627–652.
Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement errors. *Journal of Marketing Research, 18*(1), 39–50.
Fragapane, G., Ivanov, D., Peron, M., Sgarbossa, F., & Strandhagen, J. O. Increasing flexibility and productivity in industry 4.0 production networks with autonomous mobile robots and smart intralogistics. *Annals of Operations Research* (Forthcoming).
Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., et al. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47*, 55–130.
Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology, 3*, 405–423.
Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General, 143*, 1600–1615.
Gray, K., & Wegner, D. M. (2011). Dimensions of moral emotions. *Emotion Review, 3*(3), 258–260.
Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130.
Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124.
Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 181–217). Cambridge, MA: MIT Press.
Han, B. T., & Cook, J. S. (1998). An efficient heuristic for robot acquisition and cell formation. *Annals of Operations Research, 77*, 229–252.
Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011a). Can you trust your robot? *Ergonomics in Design, 19*(3), 24–29.
Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011b). A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors, 53*(5), 517–527.
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.
IBM. (2019). Building trust in AI. Retrieved 3, December from 2019. https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html.
Kats, V., & Levner, E. (1997). Minimizing the number of robots to meet a given cyclic schedule. *Annals of Operations Research, 69*, 209–226.
Kim, D. J., Song, Y. I., Braynov, S. B., & Rao, H. R. (2005). A multidimensional trust formation model in B-to-C e-commerce: A conceptual framework and content analyses of academia/practitioner perspectives. *Decision Support System, 40*(2), 143–165.
Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly, 30*(4), 941–960.
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*(10), 880–918.
Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data and Society, 5*(1), 1–16.
Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems, 17*(1), 39–71.
MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*, 297–337.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.
McAndrew, F. T., & Koehnke, S. S. (2016). On the nature of creepiness. *New Ideas in Psychology, 43*, 10–15.
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 (TMIS), 2*(2), 1–25.
Mori, M. (1970). The uncanny valley. *Energy, 7*(4), 33–35.
Nissen, M. E., & Sengupta, K. (2006). Incorporating software agents into supply chains: Experimental investigation with a procurement task. *MIS Quarterly, 301*(1), 145–166.
Pavlou, P. A., & Gefen, D. (2005). Psychological contract violation in online marketplaces: Antecedents, consequences, and moderating role. *Information Systems Research, 16*(4), 372–399.
Petrovic, S. (2019). "You have to get wet to learn how to swim" applied to bridging the gap between research into personnel scheduling and its implementation in practice. *Annals of Operations Research, 275*(1), 161–179.
Petter, S., Straub, D. W., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4), 623–656.
Piazza, J., & Sousa, P. (2016). When injustice is at stake, moral judgements are not parochial. *Proceedings from the Royal Society of London B, 283*, 20152037.
Piazza, J., Sousa, P., Rottman, J., & Syropoulos, S. (2019). Which appraisals are foundational to moral judgment? Harm, injustice, and beyond. *Social Psychological and Personality Science, 10*(7), 903–913.
Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A Critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.
PwC. (2017). Sizing the price: What's the real value of AI for your business and how can you capitalize. White Paper, Retrieved 2, December from 2019. https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html.
Ringle, C. M., Wende, S., & Becker, J.-M. (2015). *SmartPLS 3*. Boenningstedt: SmartPLS GmbH.
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*(3), 393–404.
Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion, 11*, 233–240.
Saif, I., & Ammanath, B. (2020). 'Trustworthy AI' is a framework to help manage unique risk. *MIT Technology Review* (March).
Seeber, I., Waizenegger, L., Seidel, S., Morana, S., Benbasat, I., & Lowry, P. B. Collaborating with Technology-Based Autonomous Agents: Issues and Research Opportunities. *Internet Research* (Forthcoming).
Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*(September), 401–411.
Shank, D. B., Graves, C., Gott, A., Gamez, P., & Rodriguez, S. (2019). Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computer in Human Behavior, 98*, 256–266.
Siemens. (2019). Artificial intelligence in industry: Intelligent Production. Retrieved 2, December from 2019. https://new.siemens.com/global/en/company/stories/industry/ai-in-industries.html.
Sousa, P., & Piazza, J. (2014). Harmful transgressions qua moral transgressions: A deflationary view. *Thinking and Reasoning, 20*(1), 99–128.
Srivastava, S. C., & Chandra, S. (2018). Social presence in virtual world collaboration: An uncertainty reduction perspective using a mixed methods approach. *MIS Quarterly, 42*(3), 779–803.
Talbi, E.-G. (2016). Combining metaheuristics with mathematical programming, constraint programming and machine learning. *Annals of Operations Research, 240*(1), 171–215.
Umphress, E. E., Simmons, A. L., Folger, R., Ren, R., & Bobocel, R. (2013). Observer reactions to interpersonal injustice: The roles of perpetrator intent and victim perception. *Journal of Organizational Behavior, 34*(3), 327–349.
Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology, 12*(3), 243–250.
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*(May), 113–117.
Zhong, C. B., & Leonardelli, G. J. (2008). Cold and lonely: Does social exclusion literally feel cold? *Psychological Science, 19*(9), 838–842.
Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. New York, NY: Basic Books.