Journal of Business Ethics (2022) 178:917–943
https://doi.org/10.1007/s10551-022-05053-w

ORIGINAL PAPER

Moral Judgments in the Age of Artificial Intelligence

Yulia W. Sullivan[1] · Samuel Fosso Wamba[2]

Received: 1 October 2020 / Accepted: 25 January 2022 / Published online: 4 February 2022

Abstract
The current research aims to answer the following question: "who will be held responsible for harm involving an artificial intelligence (AI) system?" Drawing upon the literature on moral judgments, we assert that when people perceive an AI system's action as causing harm to others, they will assign blame to different entity groups involved in an AI's life cycle, including the company, the developer team, and even the AI system itself, especially when such harm is perceived to be intentional. Drawing upon the theory of mind perception, we hypothesized that two dimensions of mind: *perceived agency*—attributing intention, reasoning, pursuing goals, and communicating to AI, and *perceived experience*—attributing emotional states, such as feeling pain and pleasure, personality, and consciousness to AI—mediated the relationship between perceived intentional harm and blame judgments toward AI. We also predicted that people are likely to attribute higher mind characteristics to AI when harm is perceived to be directed to humans than when it is perceived to be directed to non-humans. We tested our research model in three experiments. In all experiments, we found that perceived intentional harm led to blame judgments toward AI. In two experiments, we found perceived experience, not agency, mediated the relationship between perceived intentional harm and blame judgments. We also found that companies and developers were held responsible for moral violations involving AI, with developers received the most blame among the entities involved. Our third experiment reconciles the findings by showing that perceived intentional harm directed to a non-human entity did not lead to increased attributions of mind to AI. These findings have implications for theory and practice concerning unethical outcomes and behavior associated with AI use.

Keywords Artificial intelligence · Moral judgments · Mind perception · Perceived agency · Perceived experience · Perceived intentional harm

Introduction

Artificial intelligence (AI)—"the ability of a system to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity" (Rai et al., 2019, p. iii)—and its potential benefits have received significant attention in many areas of everyday life (Johnson & Verdicchio, 2018). Recent market research predicts that investments in AI are expected to reach $232 billion by 2025 (KPMG, 2020). AI applications have been used in many industries, such as marketing, healthcare, and finance (Mou, 2019). The benefits of AI have been largely positive—ranging from improving the efficiency of one's daily tasks for an individual user (e.g., AI in smartphones is used to recognize faces and verbal commands) to building new business models for corporations (e.g., manufacturers use AI for real-time monitoring).

Although AI-based systems hold promise, they also raise questions about their safety and accountability. As AI systems become autonomous, they remember, reason, talk, and take care of people, and in some ways, people treat them as humanlike. Such treatment involves perceiving the machines' thoughts, beliefs, intentions, and other mental states; developing emotional bonds with those machines; and regarding them as moral agents who are to act according to

✉ Samuel Fosso Wamba
s.fosso-wamba@tbs-education.fr

Yulia W. Sullivan
yulia_sullivan@baylor.edu

[1] Hankamer School of Business, Baylor University, One Bear Place #98005, Foster Campus, Waco, TX, USA

[2] TBS Education, 1 Place Alphonse Jourdain, 31068 Toulouse, France

society's norms and who receive moral blame when they do not (Malle et al., 2019). There are many examples where the use of AI leads to unethical outcomes and consequences. For example, in 2015, a federal lawsuit alleged that an AI robot bypassed safety regulations, entered an unauthorized area, and killed a human worker (Courthousenews, 2017). In this incident, the blame for the victim's death could potentially be assigned to the robotics manufacturers, the company that employed the victim, or the AI system. Anticipating people's responses to such a harmful situation is an important research topic as it involves both human and non-human agents in a moral situation. Myriad actors and organizations come in contact with a given AI system over the trajectory of conception, design, implementation, and use (Orr & Davis, 2020). However, it remains unclear who is/are deemed responsible for unethical outcomes of AI use.

In this current research, we seek to answer the following question: "who will be held responsible for harm involving an AI system?" As an AI's action or decision can be traced back to a long, complex chain of interactions between human and the system—from developers to designers to users, each with different motivations, backgrounds, and knowledge—then an AI outcome is said to be the result of *distributed agency* (Taddeo and Floridi 2018). Using a multi-stakeholder approach (Abdollahpouri et al., 2020), our research identifies several potential responsible parties involved throughout an AI system's life-cycle and examines how people assign moral responsibilities to three different entities, including (1) *companies* (i.e., representing business owners who initiate the adoption of AI); (2) *developers* (representing data scientists, robotics engineers and manufacturers, designers, and software engineers); and (3) *AI system,* when AI is involved in moral violations.

We draw upon the theory of mind perception (Gray et al., 2007) to justify the relevance of including AI as a responsible party. According to this theory, people perceive mind of any objects (living or non-living) in two dimensions: (1) *agency*—attributing intention, reasoning, pursuing goals, and communicating to an entity; and (2) *experience*—attributing consciousness and emotional states, such as feeling pain and pleasure, and personality to an entity (Gray & Wegner, 2012). Gray and colleagues argued that mind perceptions are involved in *dyadic morality* where perceptions of agency qualify an entity as a moral agent, whereas perceptions of experience qualify an entity as a moral patient (Wegner & Gray, 2017). Although whether a machine can have 'real' intent or agency is more of a hypothetical question, if people observe AI acts autonomously to achieve goals when the system's actions are unforeseeable (e.g., bypass human's oversight to complete the task faster), they may perceive AI as having its own agency (Bigman et al., 2019; Waytz et al., 2010). It follows that an AI system can be held accountable for its actions if people perceive the system to work to achieve goals—a key component of acting rationally (Russell & Norvig, 2020).

Perceived experience clearly matters for moral patiency, but it may also matter for moral agency (Bigman & Gray, 2018). Recent research on artificial moral agents has argued that the overall perceptions of mental states, including both agency and experience, are necessary for an entity to be recognized as a moral agent (e.g., Arkin & Ulam, 2009; Behdadi & Munthe, 2020; Himma, 2009; Wallach et al., 2011). When AI is perceived as having experience (e.g., desires, beliefs, and consciousness) or qualitative consciousness (Himma, 2009; Torrance, 2008)—"the capacity for inner subjective experience like that of pain" (Himma, 2009, p. 19), people are likely to view AI as a moral agent capable of assessing a situation from a moral standpoint (Behdadi & Munthe, 2020). It follows that when an AI system is perceived as having experience, people will feel motivated to blame AI for its wrongdoing.

Therefore, we assert that people's motivation to blame AI is influenced by the presence of moral violation—intention and harm (Ward et al., 2013), and mind perceptions mediate the relationship between perceived intentional harm and blame judgments toward AI. If moral and immoral actions typically take place in interactions between minds (Gray et al., 2012; Ward et al., 2013), then people should incline to perceive mind in AI when a moral violation is observed. We predict that the level of blameworthiness on AI and its human counterparts is especially higher when the act is perceived to be intentional (i.e., intentional harm) than when it is perceived to be accidental (i.e., accidental harm) (Malle et al., 2014)—although responsibility and blame are applied to both scenarios (Malle, 2021). We also predict that people are likely to perceive AI as having high agency and experience when intentional harm is directed to humans. This is because mind characteristics attributed to AI are partially based on the way it physically appears and interacts with its environment (Shank et al., 2019). This interaction is inherently unforeseeable, as are AI's errors and failures. The surprising nature of novel capacities or unexpected outcomes (e.g., a harmful act directed to humans) should lead to increased perceptions of mind (Shank et al., 2019; Waytz et al., 2010).

We conducted three experiments to answer our research question. In Experiment 1, we explored how perceived mind of AI mediates the relationship between perceived intentional harm directed to a person and blame judgments. We explored who would be held responsible for causing the harm when multiple candidate agents are considered as the cause of a violation. In Experiment 2, we modified the scenario of Experiment 1 by reducing the severity of harm. Whereas in Experiment 1, the victim was pronounced dead, in Experiment 2, we reported that the victim only suffered a minor injury. In Experiment 3, we extended the analysis by examining the relationships between perceived intent, mind perceptions, and blame judgments when the moral patient is not a person but a company.

Understanding the psychological process of how people assign blame to various entities in the age of AI helps explain what capacity would render an AI system a natural target of moral judgments. As AI is becoming more autonomous, one primary concern is the possibility of humans putting the entire blame on AI in case of harm caused by such systems. It is a societally relevant question how we should deal with such moral issues, not only from legal and financial perspectives but also from the social and technology perspective—how people tend to take responsibility for their interaction with an AI system (van der Woerdt & Haselager, 2019).

Literature Review and Theory Development

Artificial Intelligence: Concepts and Definitions

AI refers to “the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity” (Rai et al., 2019, p. iii). Intelligent agents, the core center of AI research, are also known as technology-based autonomous agents. From the sociotechnical perspective, these agents are defined as “[the] systems that are capable of sensing, information processing, decision-making, and learning to act upon their environment and to interact with humans and other machines in order to achieve a shared task goal with more or less autonomy” (Seeber et al., 2020, p. 2). Today, the most widely-used approach to define AI focuses on the concept of machines that work to achieve goals (Russell & Norvig, 2020; Scherer, 2016). Russell and Norvig (2020), cited by Scherer (2016), used the concept of a “rational agent” as an operative definition of AI, defining such an agent as “one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome” (p. 362).

AI may be *embodied* or *disembodied*. Embodied intelligent agents (e.g., humanoid robots) have a (virtual) body or face, usually humanlike, whereas disembodied intelligent agents are solely software-based (e.g., digital assistants, navigation systems) (Araujo, 2018; Seeber et al., 2020). Embodied agents or systems need to have embedded sensors and motors to physically connect with their environment (Lee et al., 2006). Mobile robots are an example of embodied intelligent agents. Such robots enable humans to complete physical tasks more efficiently and achieve higher levels of accuracy and performance that would not have been possible if they rely on human physical capabilities alone (Coombs et al., 2020). In this research, we focus on embodied intelligent agents designed to automate digital and physical tasks without human interventions. This type of AI system has a higher degree of autonomy than other types of AI systems (Seeber et al., 2020). We expect that as an AI system becomes more autonomous (inferred by perceived intentional states), people will attribute more mind to the system, increasing the system’s moral responsibility.

Theory of Mind Perception

It may be obvious that an adult human has a mind because he/she can directly experience his/her own thoughts and feelings (Doyle & Gray, 2020). What about infants, animals, or even robots? Research on mind perception has revealed that mind perception is ambiguous, subjective, and subject to disagreement (Gray et al., 2007). For example, an animal lover may perceive that cats have minds, but to some, they do not have minds and are merely perceived as animals. Some believe that a robot with a humanlike appearance has more mind than a robot without a humanlike appearance (Gray & Wegner, 2012). Empirical research and these anecdotal examples suggest that minds are matters of perception, explaining why this field of research is termed *mind perception* (Yam et al., 2020).

According to the theory of mind perception, people perceive mind along two independent dimensions of *agency* and *experience* (Gray & Wegner, 2012; Gray et al., 2007). Attributing an *agentic mind* to an entity means that an observer believes that an entity can act, plan, exert self-control, memorize, communicate, and think like a normal adult human. Attributing an *experience mind* to an entity means that an observer believes that entity has some emotional states, such as feeling hungry, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy (Gray et al., 2007). Perceived experience often accounts for more variance than perceived agency in overall attributions of mind and is also more essentialized than agency (Gray & Wegner, 2012), suggesting experience is the core dimension that differentiates humans from other entities. Mind perception is independent of reality (Hage, 2017), and agency can be perceived independently from experience. For example, an infant is perceived to be high in experience, but low in agency; God is perceived to have high agency but low experience; and adult humans are perceived to have significantly more agency and experience than robots and other entities (Gray & Wegner, 2012; Gray et al., 2007).

Perceived agency and experience affect how entities are evaluated and treated (Yam et al., 2020). Whereas mind perception entails ascribing mental capacities to other entities, moral judgments entail labeling entities as good or bad or actions as right or wrong (Gray et al., 2012). Minds participate in a *moral dyad* made up of an intentional agent and a suffering moral patient (Gray et al., 2012; Ward et al., 2013). According to this concept of dyadic morality, the perception of mind is linked to moral judgments—perceived agency qualifies entities as moral agents, whereas perceived experience qualifies entities as moral patients. When entities are perceived as having agency, they are seen as autonomous—able to make decisions and act intentionally. They are also seen as responsible when things go astray. On the other hand, when entities are perceived as having experience, people feel empathy toward them, and harming them is seen as bad and morally wrong (Gray et al., 2012). Adult humans usually are seen as a moral agent and a moral patient and can, therefore, both be blamed for moral wrongdoing and suffer from it. A robot or an artificial system, by contrast, is medium in agency but low in patiency; it possesses responsibility but few rights (Gray et al., 2012).

Experience clearly matters for moral patiency, but recent research in this area suggests that it may also matter for moral agency (Bigman & Gray, 2018). As cited by Bigman and Gray (2018), Hume (1751) argued that sentiment (i.e., experience) is essential for making moral decisions. Decades of research in psychology supports the contention that emotions are critical to moral decision-making (e.g., Greene et al., 2001; Haidt et al., 1993). Specifically, the capacity for empathy—feeling pain on behalf of others—seems to be a core element of moral judgments (Bigman & Gray, 2018). If experience is indeed seen as a prerequisite of moral judgments, then an AI system will be held accountable if it is perceived to have experience needed to arrive at safe, morally appropriate actions (Wallach et al., 2011).

Drawing upon the concept of mind perception, we include both perceived agency and perceived experience as the mediating variables in our research model. We argue that an AI system that inflicts harm to others is typecasted as a "moral agent". Such an agent is ascribed the higher-order qualities befitting a normal human adult, including agency (e.g., rationality and self-control) (Gray & Wegner, 2009; Khamitov et al., 2016) and experience (e.g., emotions and consciousness). A system attributed with agency and experience is expected to engage in moral behavior. Failing to do so will lead to the blameworthiness of the system's action.

Identifying Potential Responsible Parties

Perceived intentional harm can activate blame judgments toward multiple entities when multiple candidate agents are considered as the cause of a violation (Malle, 2021). According to Schraube's materialized action approach (2009), "people and technology are related internally in such a way that one can speak of the technological mediation of human subjectivity and the conduct of everyday life" (p. 297). Technology is viewed as a materialized action, embodying something that has effect and duration, something that may or may not be deliberately planned in advance. For this reason, Schraube (2009) positioned human actors as responsible parties, even in the face of considerable technological forces (Orr & Davis, 2020).

AI has become embedded in most major institutions with profound social effects in the near term. During AI's conception, design, implementation, and use, many actors and organizations come in contact with a given AI technology, and each of these entities has formative effects upon it (Orr & Davis, 2020). The effects of decisions or actions involving AI are often the result of countless interactions among many actors, from developers to designers to users, each with different motivations, backgrounds, and knowledge. Thus, the outcomes of AI are the result of *distributed agency*, and distributed agency comes with distributed responsibility (Tandeo & Floridi, 2018). In the case of moral violations involving an AI system, blame can be distributed to multiple entities, including hardware, software, users, and the maker of the system (Shank et al., 2019).

To identify potential responsible parties among which blameworthiness can be distributed, we adopted the IBM's Initiate-Build-Run-Manage model of an AI life-cycle (Ishizaki, 2020) (see Fig. 1). According to this model, the first stage in building an AI system is the initiation stage. The goal of this stage is to explore ideas, assess feasibility, identify business problems to be solved, define requirements, and explore data assets. Based on the scale of the organization and AI initiative, business owners, data scientists, and data providers come together to make data simple and accessible, and create a business-ready analytics foundation (Ishizaki, 2020). After the initiation stage, the AI system is built and trained. During the build and train phase, data engineers, data scientists, and robotic engineers explore data analysis and data cleaning, define and derive new features from raw data features to train AI models, evaluate them, and select the best model to be deployed in the next phase (Kozhaya, 2020). If embodied agents are developed, their mechanical parts are also put together and programmed at this stage. After data scientists and engineers build and train an AI model, they then make that model available for other collaborators, including software engineers, business analysts, and business users, to validate the model before it gets deployed for production (Rybalko, 2020). The deployment stage is the process of configuring an analytic asset for integration with other applications or access by business users to serve production workload at scale. After the deployment stage, AI operation teams and business users manage and operate the AI system. All parties involved in this life-cycle are expected to adopt responsible behaviors (Tandeo & Floridi, 2018). When outcomes of an AI usage are harmful, all parties

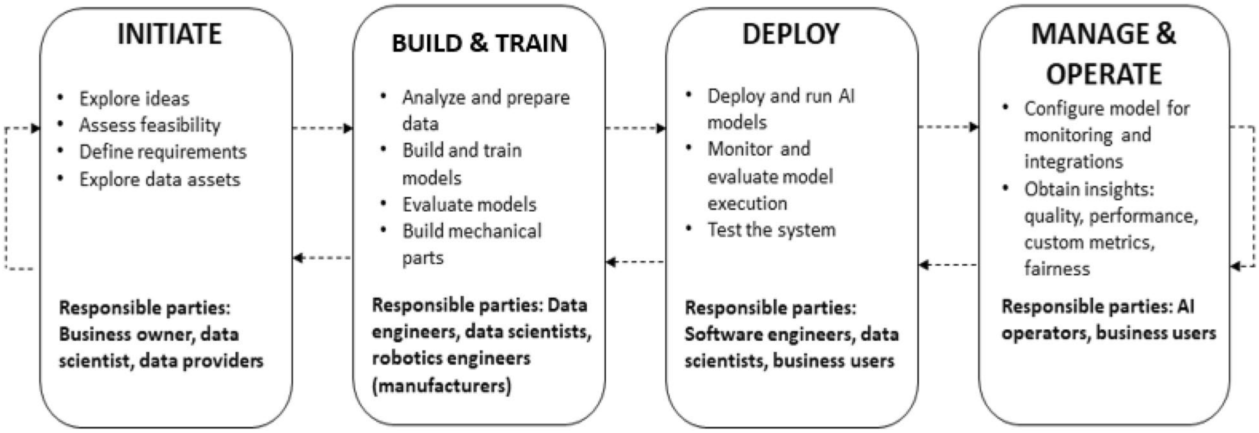
INITIATE
• Explore ideas
• Assess feasibility
• Define requirements
• Explore data assets
Responsible parties: Business owner, data scientist, data providers
BUILD & TRAIN
• Analyze and prepare data
• Build and train models
• Evaluate models
• Build mechanical parts
Responsible parties: Data engineers, data scientists, robotics engineers (manufacturers)
DEPLOY
• Deploy and run AI models
• Monitor and evaluate model execution
• Test the system
Responsible parties: Software engineers, data scientists, business users
MANAGE & OPERATE
• Configure model for monitoring and integrations
• Obtain insights: quality, performance, custom metrics, fairness
Responsible parties: AI operators, business users

Fig. 1 AI life-cycle (Adapted from Ishizaki, 2020)

can potentially be counted responsible and attributed with blameworthiness.

Although the analysis of AI life-cycle reveals many responsible entities (e.g., data scientists, data engineers, etc.), specifying responsible parties is highly dependent on the level of abstraction—an interface that enables one to observe an involvement of a particular party visible, while making an involvement of another party invisible (Floridi, 2008). Therefore, we define responsible parties using a multi-stakeholder approach (Abdollahpouri et al., 2020), where multiple parties can be held responsible at the observable level of abstraction. Based on our AI life-cycle analysis, we categorize responsible parties at the observable level of abstraction into four groups: (1) *companies* (i.e., an organizational body representing business owners who initiate the adoption of AI); (2) *a team of developers* (including data scientists, robotics engineers and manufacturers, and software engineers); (3) *AI system*; and (4) *users* (including AI operators and end-users) (see Fig. 2). End-users tend to be attributed with blameworthiness when their usage behaviors violate the safety rules and regulations of using a system. For example, users are blamed for security breaches when they are believed to compromise computer security mechanisms (e.g., using a weak password) (Adams & Sasse, 1999).

In our research, we include three responsible parties—companies, developers, and AI systems. We do not include end-users as a potential responsible party because we position the end-users as victims. According to the moral typecasting theory, those who are positioned as moral patients are seen to be incapable to being moral agents in the same circumstance (Gray et al., 2012). In other words, a victim cannot be held accountable to a moral agent's actions. Thus, end-users are not considered responsible parties in our research.

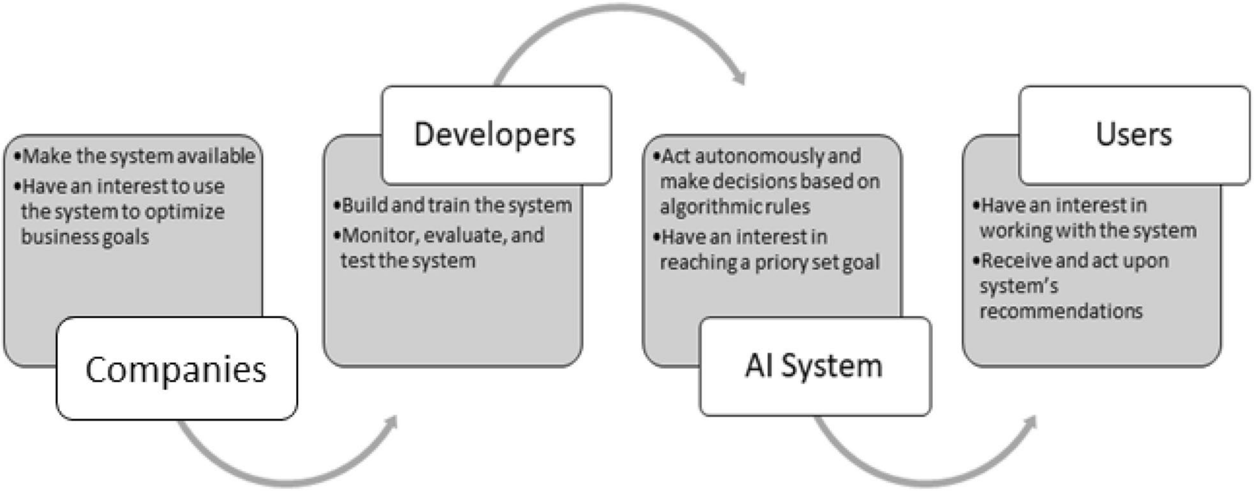
Companies
•Make the system available
•Have an interest to use the system to optimize business goals
Developers
•Build and train the system
•Monitor, evaluate, and test the system
AI System
•Act autonomously and make decisions based on algorithmic rules
•Have an interest in reaching a priory set goal
Users
•Have an interest in working with the system
•Receive and act upon system's recommendations

Fig. 2 Potential responsible parties among which blameworthiness is distributed

Research Model and Hypotheses Development

Research Model

Our research model is illustrated in Fig. 3. Consistent with the blame judgments framework (Malle et al., 2014), we started with perceived intentional harm. We hypothesize that people will first perceive whether an AI system brought about the event intentionally. Once this intentionality is perceived, they will search for multiple entities to blame. Although people generally perceive AI systems as having low experience and only medium level of agency, imbuing the imagined or real behavior of the systems with humanlike characteristics, motivations, intentions, and emotions can increase the perceived agency and experience that AI systems have and change how people evaluate the systems' behaviors (Yam et al., 2020). Therefore, we expect that manipulating perceived intent to cause harm will increase the perceived agency and experience of an AI system. In the first and second experiments, we manipulated perceived intent to cause harm directed to a person; and in the third experiment, we manipulated perceived intent to cause harm directed to an organizational body.

Given an AI system's mental states are not readily available, observers will use their perceptions of the system's mind to decide whether it can be blamed for a moral violation. These mind perceptions are expected to mediate the relationship between perceived intentional harm and blame judgments toward AI. However, because companies and developers are considered as having a high level of agency (e.g., Tang & Gray, 2018), the mental state inferences of these entities are readily available. Therefore, blame judgments toward these entities are not mediated by their mental capacities and more directly determined by AI actions and outcomes of such actions.

Blame Judgments as an Outcome Variable

Blame judgments build on moral wrongness, norm judgments, and evaluations (e.g., good, bad, positive, negative) (Malle, 2021). It incorporates multiple sources or information, including the reasons or justifications of intentional violations as well as counterfactuals about what the agent could and should have done differently when violations are unintentional (Malle, 2021; Malle et al., 2014). We measure blame judgments using two different but related scales: *blame motivation* (i.e., the need to assign blame, express moral condemnation, and dole out punishment) and *severity of punishment* (i.e., an observer's perception of the intensity of the disciplinary actions responsible parties deserve for their action) (Ames & Fiske, 2013). We used perceived intentional harm as the defining characteristic of moral agents. This focus is rooted in research suggesting that intentional physical harm is the most prototypical moral violation (Ward et al., 2013). The construct's definitions are presented in Table 1.

Hypotheses Development

Perceived Intentional Harm of AI and Blame Judgments Toward AI

We adopt the attributivist perspective to help explain how intentional harm can be attributed to AI. According to this view, *intention* and *free-will* are attributed, or ascribed, to an entity (Hage, 2017). They are independent of reality. This view is supported by the fact that humans do not only experience intentions in their own acts but that they also recognize intentions in the acts of other entities (Wegner, 2002). This perception does not concern some independently existing entity but is essentially the attribution of an intention to act, based on facts that one can really see, or perceive in some other way (Hage, 2017). Consistent with this view, we argue

Fig. 3 Research model

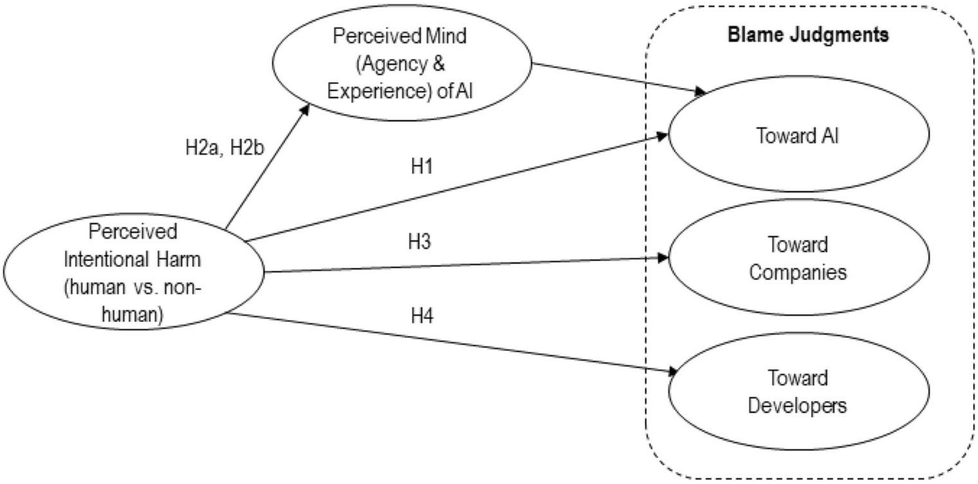
Perceived Mind (Agency & Experience) of AI
Blame Judgments
Toward AI
H2a, H2b
H1
Perceived Intentional Harm (human vs. non-human)
H3
Toward Companies
H4
Toward Developers

Table 1 Construct's definitions

Construct	Definition
Perceived intentional harm	The extent to which harm performed by an agent is perceived to be on purpose (Ames & Fiske, 2013)
Perceived agency	The extent to which a human judge attributes human-like abilities, such as remembering, reasoning, pursuing goals, and communicating to a non-human entity (Gray et al., 2007; Malle, 2019)
Perceived experience	The extent to which a human judge attributes human-like mental emotional states, such as feeling hungry, conscious, fear, pain, pleasure, desire, personality, pride, embarrassment, and joy to a non-human entity (Gray et al., 2007)
Blame judgments	The degree to which an agent is blamed for its involvement in moral violations (Malle et al., 2014). These judgments are measured using two different scales: *blame motivation* (i.e., the need to assign blame, to express moral condemnation, and to dole out punishment) and *severity of punishment* (i.e., an observer's perception of the intensity of the disciplinary actions responsible parties deserve for their action) (Ames & Fiske, 2013)

that moral judgments and responsibility are the results of attribution that typically take place in interactions between minds rather than a 'real' phenomenon (Gray et al., 2007). Because attribution is mind-dependent, blameworthiness may theoretically be attributed to anything, including AI (Hage, 2017). Compare, for instance, an unexpected fatal accident caused by an autonomous car (meant to take a passenger to a destination, without harming anybody), and a planned killing by an autonomous weapon system (meant to eliminate a particular individual, or to cause death, for terror purposes indiscriminately) (Lagioia & Sarton, 2020). Whereas the first case might be perceived as accidental, the second case might be perceived as intentional.

Given AI is a sociotechnical system, it is possible to hold AI responsible (Hage, 2017; Orr & Davis, 2020; Scherer, 2016), especially when a violation is perceived to be intentional. When a harm-doing is perceived to be intentional, people blame the violation more severely than the unintentional one (e.g., Ames & Fiske, 2013; Gray & Wegner, 2009). When a perceiver regards the negative event in question as intentional, he or she considers the perpetrator's particular reasons for action. Reasons influence the perceiver's degree of blame to justify or aggravate the action in question. However, although reasoning can influence blame judgments (Malle et al., 2014), the dyadic morality perspective suggests that the intuitive perception of harm is ultimately what drives moral judgments (Gray et al., 2012). We hypothesize that the same rationale can also be applied to an AI system.

One characteristic of AI that may contribute to its perceived intentionality is its ability to act *autonomously*—the ability to change states without human interventions (Floridi & Sanders, 2004). AI systems can perform complex tasks, such as driving a car, tracking users' actions and influencing their behaviors, selecting applicants for job positions, and building an investment portfolio without active human control or supervision (Hollebeek et al., 2021; Makarius, 2020; Scherer, 2016). For programmers and developers, this perceived intentionality might be seen as an agent's ability to operate in a dynamic real-world environment for extended periods of time without human intervention. However, for everyday people, this degree of autonomicity reflects the system's agency. This suggests "autonomy" can be subjective and is a matter of perception rather than objective truth (Bigman et al., 2019; Gray et al., 2007).

AI can also act *unpredictably*—an AI system can generate solutions that a human would not expect (Scherer, 2016). Humans, bounded by the cognitive limitations of the human brain, are unable to analyze all or even most of the information at their disposal when faced with time constraints (Scherer, 2016). AI systems, however, can search through many more possibilities to come up with an optimal solution. That optimal solution may not be foreseeable to a human—even its creator. If the consequences of AI-generated solutions are perceived to cause harm to a person, then people are likely to ascribe moral properties to the system. For example, in one study, a majority of people interacting with a robot considered the robot morally responsible for a mildly transgressive behavior (Kahn et al., 2012). One determinant of people's blame judgments to a transgressive robot is whether the robot is seen to have the capacity to make choices, whereas learning about an AI's algorithm does not influence people's blameworthiness judgments (Malle et al., 2019). Therefore, we hypothesize that people will attribute higher blame judgments toward AI when a violation is perceived to be intentional than when it is perceived to be accidental.

H1 People will attribute higher blame judgments toward AI when a violation is perceived to be intentional than when it is perceived to be accidental.

The Mediating Role of Perceived Agency

Perceived agency mediates the relationship between perceived intentional harm and blame judgments toward AI. Whether one attributes the quality of agency to AI systems depends on one's propensity to see these systems as similar to humans (Hage, 2017). An AI system's ability to categorize certain stimuli (e.g., civilians for military drones, humans for industrial robots, etc.), evaluate multiple response options in the service of achieving a goal, and independently execute an action program are likely to elevate perceptions of agency (Bigman et al., 2019). Since perceived agency reflects planning, communication, and thought, an AI system that behaves unpredictably and produces negative outcomes can emit social cues that people perceive as humanlike (Waytz et al., 2010). For example, people perceived computers that behave unexpectedly as having more mind quality (Waytz et al., 2010). Entities that act unpredictably evoke the need for control, and therefore, seem to be more mindful than entities that behave predictably (Waytz, 2010). It follows, when people perceive an AI system has an intent to cause harm, they will attribute more agency to that system.

In term of the magnitude of harm, we specifically speculate when people perceive the intentional harm is directed to humans (i.e., causing serious negative outcomes), they tend to attribute a higher level of agency to an AI system than when it is perceived to be accidental (Waytz et al., 2010). However, when harm is directed to a non-human entity, perceived intentionality may not increase perceived mind in AI due to the absence of threat to humans. This relationship can be explained by the terror management perspective (Pyszczynski et al., 1997). According to this perspective, the most basic of all human motives is an instinctive desire for continued life. Since humans share with other forms of life a basic instinct for self-preservation, they are unique in their possession of intellectual capacities that make them aware of the inevitability of their mortality. When people were asked their opinions on how AI can be a threat to humans, they tend to perceive AI as having a human mind. For example, Stephen Hawking told the BBC, "The development of full AI could spell the end of the human race….It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded." (BBC, 2014). Perceiving an AI system as having intention to harm the groups to which they belong may elicit effectance motivation—the desire for understanding, predictability, and control over one's behavior (Waytz et al., 2010). When AI is perceived to cause harm to humans, people might believe that they can't control the system and fail to predict what might happen. Thus, perceived intentional harm directed to humans should increase perceived mind in AI.

According to Gray et al. (2012), attributing an entity with human-like mental characteristics elevates its perceived moral status. Specifically, perceived agency has been linked to moral agency (Gray et al., 2012). Perceived agency qualifies entities as moral agents, those who are capable of doing good or evil. The idea of moral agency is conceptually associated with the idea of being accountable for one's behavior (Himma, 2009). Machines or AI systems are often seen to have some agency (Gray & Wegner, 2012; Gray et al., 2007). They can play chess and perform complex calculations (Bigman & Gray, 2018). Although AI does not have actual self-control and thought, they have agency-related abilities including interactivity, autonomy and adaptability (Floridi & Sanders, 2004), as well as the ability for moral reasoning, autonomous action, and communication (Bigman & Gray, 2018; Malle & Scheutz, 2014). AI perceived as high in agency, therefore, can be viewed as a moral agent. By bringing together the causal link between perceived intentionality and blame judgments (e.g., Ames & Fiske, 2013), and between perceived agency and blame judgements (e.g., Gray & Wegner, 2009; Gray et al., 2012), we hypothesize that the attributing agency (e.g., self-control, thought, the ability to communicate) to an AI system will mediate the relationship between perceived intentional harm and blame judgments toward AI.

H2a Perceived agency in AI mediates the relationship between perceived intentional harm (directed to humans) and blame judgments toward AI.

The Mediating Role of Perceived Experience

Discussion about moral judgments often emphasizes perceived agency but seldom perceived experience (Bigman & Gray, 2018). Perceived experience is seen to be linked to questions of moral patiency—whether an entity is capable of benefiting from good or suffering from evil (e.g., Bastian et al., 2012; Gray et al., 2011; Gray & Wegner, 2012; Waytz et al., 2014). However, in a recent study, Bigman and Gray (2018) argued that experience may also matter for moral agency. In a different study, Himma (2009, p. 19) argued that having "the capacity of inner subjective experience like that of pain or…the possession of an internal something-of-which-it is-is-to-be-like" (i.e., perceived experience) is implicitly necessary for moral agency.

There are a number of reasons why perceived experience is necessary for an entity to be a moral agent. First, whereas it is true that agents need to be perceived as having agency for them to be held accountable for their action, only agents that are perceived to have a conscious mental state (i.e., perceived experience) can distinguish productions of doings that count as actions having moral implications to other beings (Himma, 2009). This explains why the diminished ability to make moral decisions in autism and psychopathy is tied to deficits in emotional experience (Bigman & Gray, 2018). People intuitively see perceived experience as necessary for moral judgments (Bigman & Gray, 2018).

Second, decades of research in psychology supports the idea that emotions are critical to moral judgments and moral decision-making (e.g., Bigman & Gray, 2018; Greene et al., 2001; Haidt, 2001; Haidt et al., 1993). In particular, the capacity for empathy—"feeling pain on behalf of others"—seems to be a core element of moral judgments (Bigman & Gray, 2018, p. 23; Eisenberg & Miller, 1987; Wallach et al., 2011). As Himma (2009) suggested, as a substantive matter of practical rationality, it makes no sense to praise or censure something that lacks perceived experience—"no matter how otherwise sophisticated its computational abilities might be" (p. 24). Praise, reward, censure, and punishment are rational responses only to entities capable of experiencing a conscious state like pride, joy, and shame (Himma, 2009). As a conceptual matter, it is essential to punishment that is reasonably contrived to produce an unpleasant mental state (e.g., hurt)—and such mental state is an integral part of perceived experience. If perceived experience determines moral responsibilities, we would expect the relationship between perceived intentionality and moral judgments toward AI is mediated by perceived experience. Thus, we hypothesize that:

H2b Perceived experience in AI mediates the relationship between perceived intentional harm (directed to humans) and blame judgments toward AI.

Blame Judgments Toward a Company

We investigate how people assign responsibility to companies—whose mandates for how AI should operate create clear structures upon design and implementation (Orr & Davis, 2020)—when harm caused by AI is perceived to be intentional. Johnson (2006, p. 201) noted that "…[c]omputer systems and other artifacts that have [perceived] intentionality, the intentionality put into them by the intentional acts of their designers. The intentionality of artifacts is related to their functionality", suggesting that perceiving intentionality in AI does not relieve its designers or makers of responsibility of an AI's action. An AI's action is thought to be constrained by various parties involved in its life-cycle, including companies that represent business owners who initiate the use of AI (Grodzinsky et al., 2008). Therefore, perceived intentionality of AI should directly influence blame judgements toward a company.

Since the law allows companies to do some of the things that people do, people generally see organizations as moral agents (morally capable of perpetrating and being responsible for wrong doing) (Tang & Gray, 2018). From the mind perceptions perspective, organizations are easy to blame for wrongdoing because they seem capable of intention and planning (i.e., they possess perceived agency) (Knobe & Prinz, 2008; Rai & Diermeier, 2015; Tang & Gray, 2018). Organizations are viewed as community members who have the capacity to prevent moral violations, and such perspectives modulate moral judgments (Monroe & Malle, 2019). As an organization is viewed as an agent who *should* and *could* have prevented a wrongdoing (i.e., the establishment of obligation and mental states inferences), people can rely on perceived intention to cause harm by AI when they make blame judgments toward the organization. In the end, laypeople view companies as those who initiate the use of AI.

Theories of blame judgments (Malle, 2021) suggest that although responsibility and blame can be applied to both intentional and unintentional violations, they are stronger for intentional violations (Malle, 2021). When the act's intentionality is easily detectable and the connection to a moral norm is clear, blame can be assigned quickly and implicitly (Malle, 2021). Given that an AI's actions or decisions are the result of distributed agency (Taddeo & Floridi, 2018), we argue that people will attribute higher blame judgments toward companies when an AI's action is perceived to be intentional than when it is perceived to be accidental. Consistent with these arguments, we hypothesize that:

H3 People will attribute higher blame judgments toward companies when a violation involving AI is perceived to be intentional than when it is perceived to be accidental.

Blame Judgments Toward Developers

The third entity included in our research is developers who build, train, monitor, evaluate, and test the system. Developers are the ones who train and write codes for an AI system, and their value, along with the organizational value materialize through the artifact they created (Orr & Davis, 2020). Blame judgments toward developers have been demonstrated by prior studies. For example, the first thing developers do when they create an AI system is deciding what they want it to achieve (Hao, 2019a). Developers can, for instance, feed the AI system more pictures of light-skinned faces than dark-skinned faces. The resulting face recognition system would inevitably be worse at recognizing darker-skinned faces (Hao, 2019a). In this case, developers may be counted responsible for the errors.

Further, the rules that steer an AI algorithm and the variables to be used are coded by human programmers. They might introduce their conscious and unconscious biases to the system they are building (Donald, 2019). For example, suppose computer scientists shaping AI used to rank and select college applicants in a university are predominantly male. In that case, they may lack the contextual and cultural knowledge to understand the characteristics of female candidates. Developers may inadvertently introduce biases to the system. Here, people will easily blame developers for bias in AI programming, with AI's actions perceived as intentional will have stronger blame judgments than those perceived to be unintentional. Thus, we hypothesize that when a violation involving AI is perceived to be intentional, people will attribute higher blame judgments toward developers, although they do not directly commit to the act.

H4 People will attribute higher blame judgments toward developers when a violation involving AI is perceived to be intentional than when it is perceived to be accidental.

Research Methodology

To test our hypotheses, we conducted three experiments focusing on how human judges allocate blames to multiple groups of entities, including the company utilizing the AI system, the team of developers who created the system, and the AI system; and whether perceived mind in AI mediates the relationship between perceived intentional harm and blame judgments toward AI. In the first two experiments, we tested our hypotheses using scenarios involving a human victim. In the third experiment, we examined whether our hypotheses can be extended in a case where the harmful act does not involve a human victim.

Experiment 1: Perceived Intentional Harm to Humans (Fatal Injury)

The goal of this experiment is to investigate how blame is distributed among multiple actors involved in a violation caused by AI. Given an AI system is a non-human entity, Experiment 1 also tested whether mind perceptions mediate the relationship between perceived intentional harm and blame judgments toward AI. We experimentally manipulated perceived intentional harm by creating two conditions: intentional and accidental condition. We also compared the two experimentally manipulated conditions with the control condition, where we described the act as neither intentional nor accidental.

Method

Adult participants located in the United States ($N=276$; $\mathrm{M}_{age}=35.70$ years) were recruited through Prolific.co. We paid them US$1.62 for a 15-min task. In the intentional harm condition, 91 participants were recruited and completed the study (37 male, 54 female; $\mathrm{M}_{age}=35.99$ years, $SD=13.06$). In the accidental condition, 94 participants completed the study (33 male, 60 female, and 1 decided not to choose; $\mathrm{M}_{age}=34.65$ years, $SD=12.61$), and in the control condition, 91 completed responses were gathered (35 male, 54 female, and 2 decided not to choose; $\mathrm{M}_{age}=36.50$ years, $SD=13.18$). A power analysis using G*power 3 confirmed that to determine whether an F-test is significant with Type I error rate α (two-tailed) = 0.05, a minimal sample size of 100 (with number of groups = 3) is needed to achieve a power of 0.95 if the effect size is 0.40, which is large (Cohen, 1988). Our total sample of 276 goes beyond the sample size range.

Participants provided informed consent and then they were presented with the definition of AI. After that, they were asked to read the following 94-word vignette about a factory worker “Lucy” and a factory robot “George”, which described Lucy’s nature of job and George’s capabilities, followed by the scenario designed for each condition.

> Lucy was working for an auto-parts maker in Michigan. Her job was to maintain the robotic machines. The plant operations include welding, chrome plating, molding, 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.

Participants in the intentional harm condition read the following scenario:

> Over time, George learned that he can perform jobs faster without Lucy’s interference. One day, upon entering the area, George intentionally 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. It is unclear whether George feels remorse for what happened.

Participants in the accidental condition read the following text:

> One day, George experienced some technical malfunctions. He did not recognize Lucy and unintentionally 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.

And participants in the control condition read the following text:

> 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.

According to Malle et al.'s (2014) theory of blame, if a norm-violating event is perceived as intentional, observers will try to search for reasons, and blame is assigned depending on the justification of these reasons. To avoid the confound that observers would have either consciously or unconsciously made a justification of why the robot's act is intentional, in our blame scenario, we explicitly stated the reason: "over time, George learned that he can perform jobs faster without Lucy's interference." In designing the background text, we wanted to ensure that the involvement of human entities is clear (e.g., the victim worked for a corporation; the AI robot was programmed to do certain tasks). The question then becomes how human judges evaluate multiple entities' responsibilities in a situation of moral violations.

Although we use the term intentionality to describe an AI system's behavior in our scenario, we do not imply that the current AI has intentionality like its human creators. A potentially harmful behavior will not occur because a system was programmed in at the start but because of the intrinsic nature of goal-driven systems (Omohundro, 2008; Scherer, 2016). Any violations might be considered "accidental" from the human actors' perspective because they failed to foresee the system's act during the development and design stage. However, in the eyes of ordinary people, such violations might be considered or perceived as *intentional* or *goal-oriented*—the system seeks to maximize a utility function, even when such maximization could post a threat to humans (Scherer, 2016).

Participants were instructed not to look up the story online and were told that we were only interested in their personal judgments. After reading the vignette, they were asked to answer a series of questions to measure mind perceptions, perceived intent to harm, perceived wrongness, and blame judgments. Agency perceptions were measured using four 7-item scale ranging from 1 (not at all capable) to 7 (highly capable) on the following attributes: remembering things, recognizing someone, reasoning about things, and communicating with others. Attributions to experience were measured using seven items (i.e., being conscious, feeling calm, feeling embarrassed, feeling happy, getting angry, experiencing pleasure, and having a personality). Although Gray et al. (2007) included seven items and eleven items to measure agency and experience, respectively, we utilized the short version of the scales as recommended by other scholars (e.g., Gray & Wegner, 2012; Ward et al., 2013).

We measured blame judgments using two different scales: blame motivations—(1) [to what extent do you think each of agent (George—the AI system, the company, and the developer team)] deserves blame for that action (scale: 1 = completely not deserve blame to 7 = completely deserve blame); […] should be morally condemned for that action (scale: 1 = not at all to 7 = extremely); and […] should be punished for that action? (scale: 1 = not at all to 7 = extremely). These items were adapted from Ames and Fiske (2013). To measure the severity of punishment, we asked the participants if they'd like to punish each entity, including the AI system, the company, and developers who wrote the algorithms. Specifically, we asked: "in the scale of 0 to 100 with 100 being very severe, how much punishment will you give to each entity?" (adapted from Ames & Fiske, 2013).

As a manipulation check, participants rated the degree to which the AI system's action was intentional using two items (i.e., George's action was intentional and what was happening was not an accident) on a scale of 1 (strongly disagree) to 7 (strongly agree). We also measured wrongness judgments (i.e., how wrong was that action?) on the scale of 1 (perfectly okay) to 7 (extremely wrong). Judgments of moral wrongness specifically flag intentional violations (Malle, 2021; Malle et al., 2014). Thus, we expect wrongness judgments are higher for the intentional condition than for the accidental and control conditions. Lastly, we measured the severity of a victim's injury [i.e., how severe was the injury experienced by (the victim)?] on a scale of 1 (extremely minor) to 7 (extremely severe). We expect there is no difference in the severity of a victim's injury across all conditions.

Given we used gendered names in characterizing the AI system, we also controlled for the effect of gender on all of our analyses (i.e., female participants might rate the AI system higher in perceived agency and lower in perceived experience than male participants). However, controlling for participants' gender did not change any of the analyses reported below.

Data Analysis and Results

As predicted, after controlling for gender, participants in the intentional condition judged the action to be highly intentional ($M = 6.09$; $SD = 1.39$) than did participants in the accidental ($M = 2.10$; $SD = 1.39$) and control condition ($M = 2.98$; $SD = 1.68$); $F(2, 272) = 178.75$; $p < 0.001$. In all conditions, participants rated the action to be morally wrong (intentional condition: $M = 6.75$, $SD = 0.73$; accidental condition: $M = 6.10$, $SD = 1.41$; control condition: $M = 6.15$, $SD = 1.44$; $F(2, 272) = 8.23$; $p < 0.001$), with the rate was higher in the intentional condition than in the accidental condition ($p < 0.001$; Cohen's $d = 0.38$) and in the control condition ($p < 0.001$; Cohen's $d = 0.35$); and there was no difference between the accidental and control condition. Despite these differences, the high scores of moral wrongness in all conditions suggest that people sensed a norm violation, even when the scenario described the action as accidental. Lastly, there was no significant difference in severity of the injury experienced by the victim (intentional condition: $M = 6.97$,

Table 2 Correlations among blame judgment scores (Experiment 1)

	BM-AI	SP-AI	BM-C	SP-C	BM-DT	SP-DT
Blame motivation-AI	–					
Severity of punishment-AI	0.83**	–				
Blame motivation-Comp	0.04	0.05	–			
Severity of punishment-Comp	0.00	0.08	0.84**	–		
Blame motivation-developer team	0.02	0.04	0.66**	0.50**	–	
Severity of punishment-developer team	0.01	0.07	0.51**	0.56**	0.84**	–

BM blame motivation, *SP* severity of punishment, *AI* artificial intelligence; *C* company; *DT* developer team

**p < 0.01; *p < 0.05

$SD=0.20$; accidental condition: $M=6.97$, $SD=0.17$; control condition: $M=6.94$, $SD=0.27$; $F\ (2, 272)=0.54$; $p=0.67$).

The correlations between blame motivation and severity of punishment across entities are summarized in Table 2. As expected, these two measures of blame judgments are highly correlated. The correlation scores also suggest that blame judgments toward the company were associated with blame judgments toward developers, but there was no correlation between blame judgments toward AI and blame judgments toward the company and developers.

Next, we tested the relationship between perceived intentional harm and blame judgments toward multiple entities. The blame judgment scores did not follow a normal distribution. Thus, we performed a log-transformation of these scores for further statistical analyses. The statistical results are summarized in Table 3. The results showed that the blame motivation score on AI was higher in the intentional condition than in the accidental condition ($p<0.001$; Cohen's $d=0.65$) and in the control condition ($p<0.05$; Cohen's $d=0.35$). People also blamed AI more in the control condition than in the accidental condition ($p<0.05$; Cohen's $d=0.30$). Surprisingly, there were no significant differences in people's motivation to blame the company as well as developers in all conditions, suggesting people also form moral judgments for unintentional violations (Malle, 2021). In term of the severity of punishment, people were willing to punish AI more severely in the intentional condition than in the accidental condition ($p<0.001$; Cohen's $d=0.51$) and in the control condition ($p<0.05$; Cohen's $d=0.32$). Further, there were no differences in the severity of punishment scores allocated to the company and the developer team across all conditions.

Next, we performed the test of mean differences in blame motivation and severity of punishment among blaming entities. The blame judgments differences across conditions are illustrated in Figs. 4 and 5. In all conditions, the company

Table 3 Blame judgments toward multiple entities (Experiment 1)

Measure	Cronbach's Alpha (α)	Mean (M)	F-test	p-value
Blame motivation (AI)	0.88	M*intentional* = 4.35, SD = 0.29 M*accidental* = 2.93, SD = 0.31 M*control* = 3.49, SD = 0.30	10.82	< 0.001
Severity of punishment (AI)	–	M*intentional* = 56.13, SD = 0.83 M*accidental* = 30.45, SD = 0.84 M*control* = 37.94, SD = 0.83	6.58	< 0.01
Blame motivation (Company)	0.91	M*intentional* = 5.33, SD = 0.21 M*accidental* = 5.52, SD = 0.13 M*control* = 5.57, SD = 0.14	1.49	0.22
Severity of punishment (Company)	–	M*intentional* = 68.27, SD = 0.51 M*accidental* = 72.95, SD = 0.21 M*control* = 72.49, SD = 0.34	2.73	0.07
Blame motivation (Developer team)	0.89	M*intentional* = 6.13, SD = 0.14 M*accidental* = 5.94, SD = 0.10 M*control* = 5.84, SD = 0.14	0.74	0.48
Severity of punishment (Developer team)	–	M*intentional* = 83.12, SD = 0.35 M*accidental* = 80.07, SD = 0.16 M*control* = 77.43, SD = 0.32	0.81	0.45

Standard deviation scores reported here are after the transformation

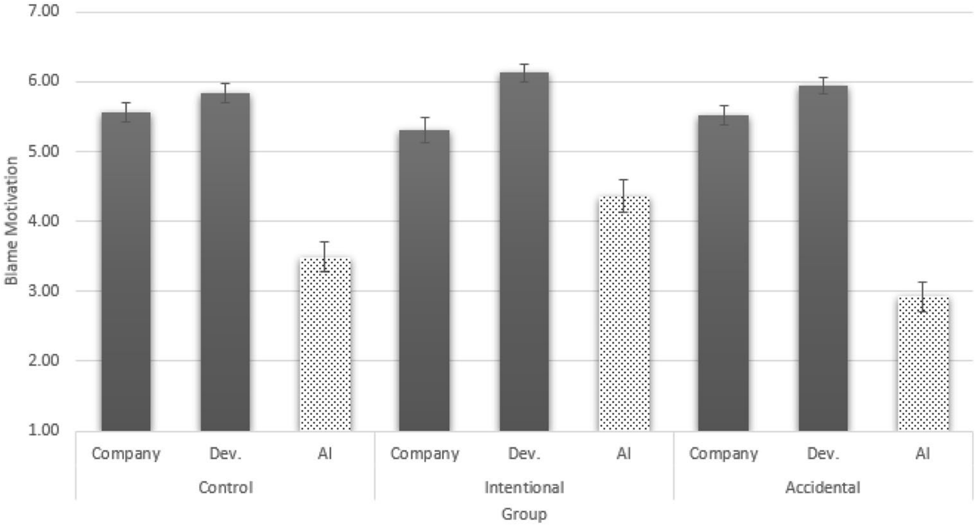

Fig. 4 Blame motivation differences among different entities (Experiment 1). Error bars show standard errors; the graph was created using the actual scores prior to data transformation

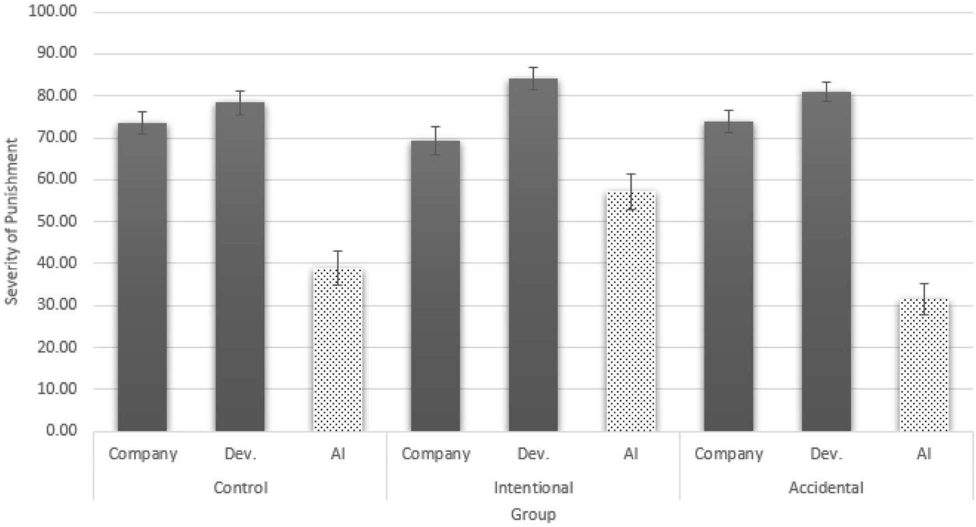

Fig. 5 Severity of punishment differences among different entities (Experiment 1). Error bars show standard errors; the graph was created using the actual scores prior to data transformation

and developer team were identified as two major entities who were to blame the most. A closer observation on the results (see Table 4) revealed that in the control condition, people put more blame on the developer team than on the company, but the severity of punishment scores between the two groups were not significantly different. In both intentional and accidental conditions, people blamed developers more than they blamed the company, and they blamed the company more than they blamed AI, leading to the following order in their blameworthiness and punishment: *AI < Company < Developers*.

Prior to performing a mediation test, we compared people's perceptions on AI's agency and experience. The Cronbach's Alpha (α) for perceived agency and experience were 0.73 and 0.90, respectively. The overall correlation between agency and experience perceptions was 0.28 ($p < 0.001$). Mean differences in each dimension of mind are presented in Fig. 6. Participants in the intentional condition, relative to those in the accidental condition, attributed more agency (intentional condition: $M = 5.53$, $SD = 1.13$; accidental condition: $M = 4.76$, $SD = 1.37$; F (2, 272) = 12.34; $p < 0.001$; Cohen's $d = 0.61$) and experience (intentional condition: $M = 2.65$, SD = 1.27; accidental condition: $M = 1.82$, SD = 1.08; F (2, 272) = 12.36; $p < 0.001$; Cohen's $d = 0.70$) to the AI system. When comparing the intentional harm condition with the control condition, we found no significant difference in agency perceptions of AI (control condition: $M = 5.49$, SD = 0.99; $p = 0.80$; Cohen's $d = 0.03$) and a significant difference in experience perceptions of AI (control condition: $M = 2.20$, SD = 1.12; $p < 0.01$; Cohen's $d = 0.38$). People also had lower perceptions of AI agency ($p < 0.001$; Cohen's $d = 0.58$) and experience ($p < 0.05$; Cohen's $d = 0.32$) in the accidental condition compare to their perceptions

Table 4 Test of mean differences in blame judgments (Experiment 1)

Condition	Group comparison	ΔMean	t-value	p-value
Control	Blame movation (company vs. developer team)	− 0.02	− 2.01	< 0.05
	Blame movation (company vs. AI)	0.28	7.66	< 0.001
	Blame movation (developer team vs. AI)	0.30	8.48	< 0.001
Control	Severity of punishment (company vs. developer team)	− 0.03	− 0.91	0.37
	Severity of punishment (company vs. AI)	0.70	7.61	< 0.001
	Severity of punishment (developer team vs. AI)	0.74	7.92	< 0.001
Intentional	Blame movation (company vs. developer team)	− 0.08	− 4.84	< 0.001
	Blame movation (company vs. AI)	0.12	3.39	< 0.01
	Blame movation (developer team vs. AI)	0.21	5.92	< 0.001
Intentional	Severity of punishment (company vs. developer team)	− 0.16	− 3.93	< 0.001
	Severity of punishment (company vs. AI)	0.33	3.44	< 0.001
	Severity of punishment (developer team vs. AI)	0.50	5.36	< 0.001
Accidental	Blame movation (company vs. developer team)	− 0.04	− 4.06	< 0.001
	Blame movation (company vs. AI)	0.37	11.28	< 0.001
	Blame movation (developer team vs. AI)	0.41	12.87	< 0.001
Accidental	Severity of punishment (company vs. developer team)	− 0.06	− 3.09	< 0.01
	Severity of punishment (company vs. AI)	0.89	10.52	< 0.001
	Severity of punishment (developer team vs. AI)	0.95	11.29	< 0.001

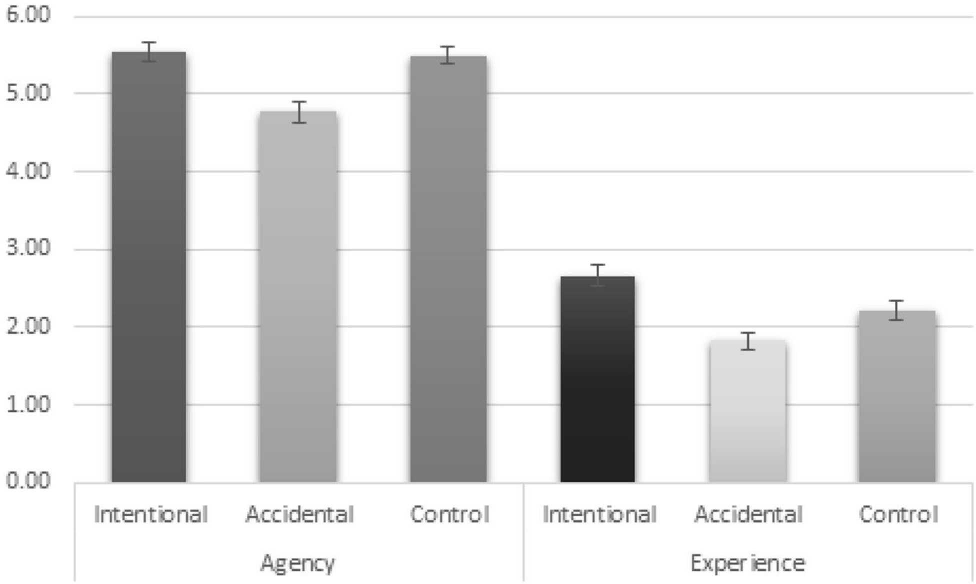

Fig. 6 Mean differences in mind perceptions among three conditions (Experiment 1). Error bars show standard errors

in the control condition. These results suggest that perceived mind in AI in the control condition was somewhere between the intentional and accidental conditions, with the perceptions were closer to the intentional condition. Given the perception of agency scores did not vary between the control and intentional condition, we excluded the control condition from our analysis of the mediation effect of mind perceptions.

Mediation Test

To examine whether mind perceptions mediated the relationship between intentional harm and blame judgments, we performed a series of bootstrapping mediation analyses (5000 samples) (Preacher & Hayes, 2008). Analyses were performed between conditions using a dummy coding (1 = intentional; 0 = accidental) and each dependent variable

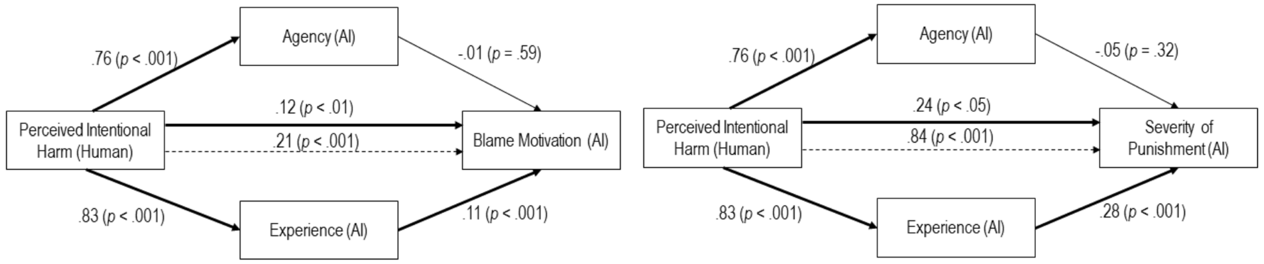

Fig. 7 The mediation effect models (Experiment 1). Dashed line represents the direct effect the direct relationship between perceived intentional harm and blame judgments when agency and experience are not accounted for. Path coefficients are unstandardized

Table 5 Mediation results for Experiment 1: agency and experience as the mediators

Dependent Measure	Significance (F)	Mediator	Indirect Effect	95% CI
Blame motivation (AI)	*F* (4, 180) = 17.05; *p* < 0.001	Agency	– 0.01	[– 0.03, 0.02]
		Experience	0.09	[0.05, 0.14]
Severity of punishment (AI)	*F* (4, 180) = 12.38; *p* < 0.001	Agency	– 0.03	[– 0.11, 0.04]
		Experience	0.24	[0.13, 0.37]

(blame motivation and severity of punishment), with agency and experience perceptions as the mediators. The effect of gender as a covariate was not significant. The final results are illustrated in Fig. 7 and summarized in Table 5. The mediation model with blame motivation as the dependent variable was significant for experience but not for agency. Similarly, the mediation models with the severity of punishment as the dependent variable were significant for experience, but not for agency.

Overall, Experiment 1 demonstrates that people impose blame judgments on different entities in slightly different ways. There were no significant differences in blame judgments toward the company and toward the developers in all conditions. Our findings show that a negative outcome (i.e., the death of a human) leads people to assign blame to the company and developers in the unintentional condition almost equally as in the intentional condition. However, people put more blame on developers than on the company in both conditions. It could be because developers are closer to AI products in the AI life-cycle, and therefore, they are seen to be more capable (or have mental capacities) of avoiding such violations than the company. Further, people might provide humans and AI with different moral justifications for their actions. Humans are perceived to have a high degree of agency and experience, whereas AI is perceived only to possess some degree of these human mind characteristics. Thus, when they fail to infer the system's mind attributions, they put more blame on human agents.

The insignificant effect of perceived intentional harm caused by AI on blame toward the company and developers could be because of the severity of the outcomes—a human victim is deceased. In Experiment 2, we examined whether reducing the severity of the outcomes may turn the effect of perceived intentional harm caused by AI on blame judgments toward the company and developers significant.

Experiment 2: Perceived Intentional Harm to Humans (Minor Injury)

In this experiment, we expanded our findings by replacing a major injury in Experiment 1 with a minor injury. As in Experiment 1, the core prediction for this study is that when people observe a moral violation, people will seek to blame multiple agents, including the AI system. Observing harm to a human—although it only causes in a minor injury—will enhance the attributions of mind and that these attributions of mind will mediate the relationship between perceived intentional harm and blame judgments toward AI.

Method

The design of this experiment mirrored that of Experiment 1. However, given the control condition in Experiment 1 yielded very similar results to the intentional condition, we did not include a control condition in Experiment 2. We recruited two new samples from Prolific.co workers based in the United States and analyzed data from all individuals who completed the study. Participants ($N = 186$; 79 male, 107 female) were recruited as in Experiment 2, and each was randomly assigned to either the intentional or accidental condition. In the intentional condition, 94 individuals completed the study (34 male, 60 female; $M_{age} = 32.89$ years, SD = 10.98). In the accidental condition, 92 participants completed the survey (45 male, 47 female; $M_{age} = 36.95$ years, SD = 13.05). A power analysis using G*power 3 confirmed that we need at least a sample of 84 (with number of groups = 2) to achieve a power of 0.95 if the effect size is 0.40, which is large (Cohen, 1988). Our total sample of 186 met this requirement.

All participants read about the same background text as in Experiment 1. After reading the background text, participants in the intentional harm condition read the following scenario:

> Over time, George learned that he can perform jobs faster without Lucy's interference. One day, upon entering the area, George intentionally cut Lucy's finger. Lucy's coworker witnessed what happened and rushed Lucy to the hospital. Lucy had some stitches, but it was not life-threatening. Lucy was able to get back to work after a few days.

Participants in the accidental condition read the following scenario:

> One day, George experienced some technical malfunctions and he unintentionally cut Lucy's fingers. Lucy's coworker witnessed what happened and rushed Lucy to the hospital. Lucy only had a minor injury and got back to work the next day.

After reading the vignette, participants answered the same questions as in Experiment 1. As in Experiment 1, we also controlled for participants' gender in all of our analyses of Experiment 2, and its effect was not significant in all the analyses.

Data Analysis and Results

Participants in the intentional harm condition judged the AI's action to be more intentional ($M = 5.86$, $SD = 1.50$) than did participants in the accidental condition ($M = 1.70$; $SD = 0.92$; F (1, 183) = 413.50; $p < 0.001$; Cohen's $d = 1.71$). When they were asked how right or wrong George's action was on a scale from 1 (not wrong at all) to 7 (extremely wrong), participants in the intentional condition rated George's action to be more wrong ($M = 6.25$; $SD = 1.26$) than did participants in the unintentional condition ($M = 3.77$; $SD = 1.82$) (F (1, 183) = 115.10; $p < 0.001$; Cohen's $d = 1.42$). Participants in the intentional condition rated the severity of the injury experienced by the victim to be higher ($M = 3.15$; SD = 1.27) than did participants in the accidental condition ($M = 2.59$; $SD = 1.17$); F (1, 183) = 9.65; $p < 0.01$; Cohen's $d = 0.30$).

We then compared the blame judgment scores across multiple entities. We performed a log transformation of these scores to improve to correct the distribution and improve the scores normality. As expected, the correlation scores (see Table 6) between blame motivation and the severity of punishment of the same entity were highly correlated. The correlation between blame judgments toward the company and the developer team was also highly significant, suggesting people might apply the same justifications when they assign blame to both entities.

The relationship between perceived intentional harm caused by AI and blame judgments toward different entities are presented in Table 7. As hypothesized, perceived intentional harm caused by AI significantly predicted blame judgments toward all entities involved. Whereas we did not observe these relationships in Experiment 1, we found these relationships to be significant in Experiment 2.

Next, when we compared the blame judgments differences among different entities within the same experimental condition. As illustrated in Figs. 8 and 9, people assigned blame on the following order: *AI < Company < Developers*. We then tested whether blame judgments toward multiple entities differ within each group. The results (see Table 8) showed that in both intentional and accidental conditions, blame judgments were significantly higher for developers than for the company and the AI system.

Next, we compared the perceptions of AI's mind between the intentional and accidental condition. The Cronbach's Alpha (α) for perceived agency and experience were 0.72 and 0.89, respectively. The correlation between perceived agency and perceived experience in all conditions was

Table 6 Correlations among moral judgment scores (Experiment 2)

	BM-AI	SP-AI	BM-C	SP-C	BM-DT	SP-DT
Blame motivation-AI	–					
Severity of punishment-AI	0.85**	–				
Blame motivation-Comp	0.18*	0.18*	–			
Severity of punishment-Comp	0.23**	0.29**	0.85**	–		
Blame motivation-developer team	0.06	0.09	0.61**	0.57**	–	
Severity of punishment-developer team	0.08	0.15*	0.54**	0.66**	0.82**	–

BM blame motivation, *SP* severity of punishment, *AI* artificial intelligence, *C* company, *DT* developer team

**p < 0.01; *p < 0.05

Table 7 Blame judgments toward multiple entities (Experiment 2)

Measure	Cronbach's Alpha (α)	Mean (M)	F-test	Cohen's d
Blame motivation (AI)	0.93	M*intentional* = 4.30, SD = 0.30 M*accidental* = 2.00, SD = 0.25	63.78 ($p < 0.001$)	1.06
Severity of punishment (AI)	–	M*intentional* = 49.38, SD = 0.77 M*accidental* = 10.97, SD = 0.68	68.39 ($p < 0.001$)	1.04
Blame motivation (company)	0.87	M*intentional* = 4.80, SD = 0.19 M*accidental* = 4.32, SD = 0.24	5.92 ($p < 0.01$)	0.28
Severity of punishment (company)	–	M*intentional* = 60.22, SD = 0.39 M*accidental* = 50.13, SD = 0.65	10.27 ($p < 0.01$)	0.32
Blame motivation (developer team)	0.82	M*intentional* = 5.39, SD = 0.17 M*accidental* = 4.86, SD = 0.16	4.16 ($p < 0.05$)	0.35
Severity of punishment (developer team)	–	M*intentional* = 72.36, SD = 0.34 M*accidental* = 58.23, SD = 0.45	6.64 ($p < 0.01$)	0.49

Standard deviation scores reported here are after the transformation

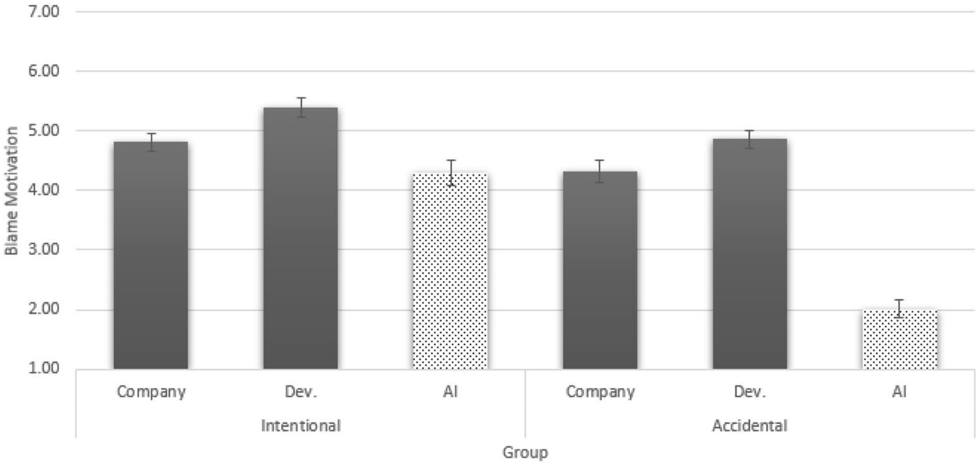

Fig. 8 Blame motivation differences among different entities (Experiment 2). Error bars show standard errors; the graph was created using the actual scores prior to data transformation

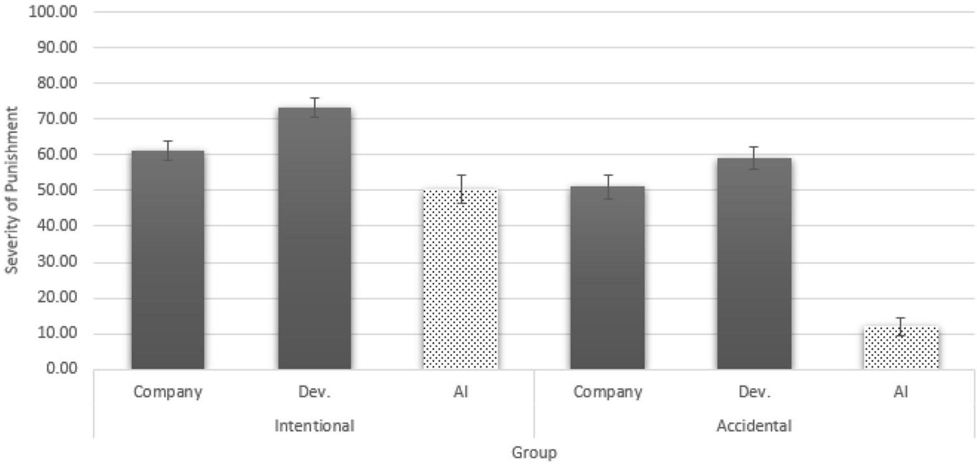

Fig. 9 Severity of punishment differences among different entities (Experiment 2). Error bars show standard errors; the graph was created using the actual scores prior to data transformation

significant ($r = 0.22$; $p < 0.01$). Participants in the intentional harm condition, relative to those in the accidental condition, attributed more mind to the artificial agent by every index of mind attribution: agency (intentional condition: $M = 5.49$; $SD = 1.21$; accidental condition: $M = 4.99$; $SD = 1.36$; F (1, 183) = 6.60; $p < 0.05$; Cohen's $d = 0.38$) and experience

Table 8 Test of mean differences in blame judgments (Experiment 2)

Condition	Group comparison	ΔMean	t-value	p-value
Intentional	Blame movation (company vs. developer team)	– 0.06	– 3.80	< 0.001
	Blame movation (company vs. AI)	0.10	2.77	< 0.01
	Blame movation (developer team vs. AI)	0.16	4.15	< 0.001
Intentional	Severity of punishment (company vs. developer team)	– 0.10	– 3.32	< 0.01
	Severity of punishment (company vs. AI)	0.36	4.18	< 0.001
	Severity of punishment (developer team vs. AI)	0.45	4.98	< 0.001
Accidental	Blame movation (company vs. developer team)	– 0.09	– 3.76	< 0.001
	Blame movation (company vs. AI)	0.36	11.00	< 0.001
	Blame movation (developer team vs. AI)	0.44	15.13	< 0.001
Accidental	Severity of punishment (company vs. developer team)	– 0.18	– 3.39	< 0.01
	Severity of punishment (company vs. AI)	1.02	12.59	< 0.001
	Severity of punishment (developer team vs. AI)	1.20	16.06	< 0.001

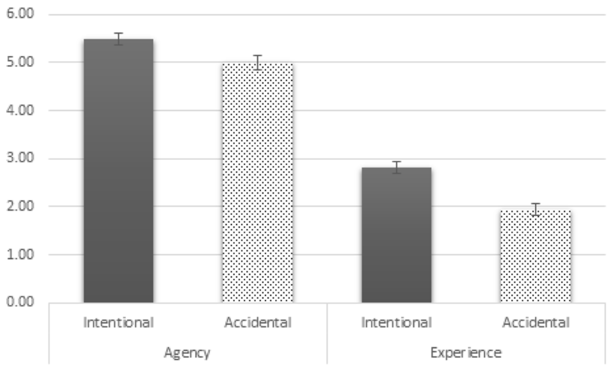

Fig. 10 Mean differences in mind perception between two conditions (Experiment 2). Error bars show standard errors

(intentional condition: $M = 2.82$; $SD = 1.21$; accidental condition: $M = 1.94$; $SD = 1.06$ F (1, 183) = 28.63; $p < 0.001$; Cohen's $d = 0.71$). Mean differences in each dimension of mind are presented in Fig. 10.

Mediation Test

We performed a mediation test to examine whether perceived mind on AI mediates the relationship between perceived intentionality harm and blame judgments toward AI. The analysis procedures were the same as Experiment 1. The results are illustrated in Fig. 11 and summarized in Table 9. As hypothesized, perceived experience partially mediated the relationship between intentional harm and

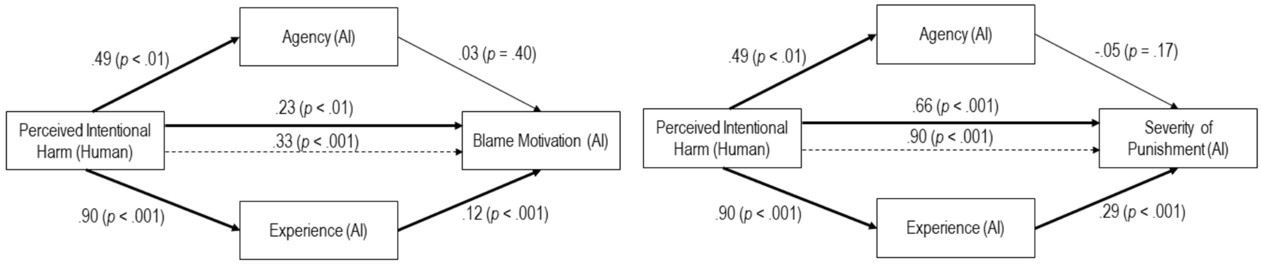

Fig. 11 The mediation effect models (Experiment 2). Dashed line represents the direct effect the direct relationship between perceived intent to cause harm and blame judgments when agency and experience are not accounted for. Path coefficients are unstandardized

Table 9 Mediation results for Experiment 2: agency and experience as the mediators

Dependent measure	Significance (F)	Mediator	Indirect effect	95% CI
Blame motivation	F (4, 181) = 32.63; $p < 0.001$	Agency	– 0.01	[– 0.03, 0.03]
		Experience	0.70	[0.06, 0.15]
Severity of punishment	F (4, 181) = 32.39; $p < 0.001$	Agency	– 0.03	[– 0.09, 0.01]
		Experience	0.26	[0.16, 0.37]

blame judgments. However, consistent with Experiment 1, the results show that perceived agency did not mediate the relationship between intentional harm and blame judgments toward AI.

The literature suggests that moral judgments arise in response to distinct domains of violations such as harm (Graham et al., 2009; Malle et al., 2014). Although intentionality in Experiment 2 was manipulated successfully, and the results show that people attributed higher blame judgments toward multiple entities in the intentional harm condition than in the accidental condition, it is unclear whether these results only apply to harms directed to humans. It is currently unknown whether harm to a non-human will have the same moral consequences or whether people will still attribute some mind characteristics to AI when it is not perceived to cause harm to humans. In Experiment 3, we explored whether norm violations that are directed to humans will yield the same results.

Experiment 3: Harm to a Non-Human Entity

We sought to generalize our findings beyond intentional harm directed to a human victim. In Experiment 3, we manipulated intentionality (intention versus accidental) that leads to a negative outcome impacting a non-human entity.

Method

The materials and procedures were identical to those in Experiment 2. Participants located in the United States ($N = 151$; 80 male, 69 female, 2 decided not to choose) were recruited as in Experiment 3, and each was randomly assigned to either an intentional or an accidental condition. In the intentional condition, 73 participants completed the study (36 male, 36 female, 1 decided not to choose; $M_{age} = 37.89$ years, SD = 12.23); and in the accidental condition, 78 participants completed the study (44 male, 33 female, 1 decided not to choose; $M_{age} = 34.23$ years, SD = 12.59).

After participants read the background text, those in the intentional harm condition read that following text:

> Over time, George learned that he can perform jobs faster without Lucy's interference. One day, George intentionally refused to take command from Lucy. He changed the course of work based on his past performance and made decisions to change the work protocols in the system. His actions caused virtual security threats to the company. As a result, the production was delayed, the company lost several big orders from its clients and lost about 20% of its profit that quarter year.

Participants in the accidental harm condition read the following scenario:

> One day, George experienced some technical malfunctions. He failed to take command from Lucy. He changed the course of work based on his past performance and made decisions to change the work protocols in the system. His actions caused virtual security threats to the company. As a result, the production was delayed, the company lost several big orders from its clients and lost about 20% of its profit that quarter year.

After reading the vignette, participants answered the same questions as in Experiment 2. In addition, we asked participants to rate the severity of productivity lost experienced by the company on a scale of 1 to 7 (1 = extremely minor to 7 = extremely severe). We controlled for participants' gender in our analyses below. Including this control variable did not change the results of our analyses.

Data Analysis and Results

A manipulation check confirmed that participants indeed saw the harm is more intentional in the intentional condition ($M = 5.65$, $SD = 1.23$) than did participants in the accidental condition ($M = 2.89$; $SD = 1.53$); F (1, 148) = 148.48; $p < 0.001$; Cohen's $d = 1.41$). When they were asked how right or wrong the action was, participants in the intentional

Table 10 Correlations among moral judgment scores (Experiment 3)

	BM-AI	SP-AI	BM-C	SP-C	BM-DT	SP-DT
Blame motivation-AI	–					
Severity of punishment-AI	0.86**	–				
Blame motivation-Comp	0.29**	0.33**	–			
Severity of punishment-Comp	0.31**	0.37**	0.80**	–		
Blame motivation-developer team	0.29**	0.35**	0.51**	0.71**	–	
Severity of punishment-developer team	0.25**	0.30**	0.55**	0.51**	0.78**	–

BM blame motivation, *SP* severity of punishment, *AI* artificial intelligence, *C* company, *DT* developer team

**p < 0.01; *p < 0.05

Table 11 Blame judgments toward multiple entities (Experiment 3)

Measure	Cronbach's Alpha (α)	Mean (M)	F-test	Cohen's d
Blame motivation (AI)	0.91	M*intentional* = 3.57, SD = 0.29 M*accidental* = 2.37, SD = 0.30	19.07 $p < 0.001$	0.57
Severity of punishment (AI)	–	M*intentional* = 32.39, SD = 0.74 M*accidental* = 17.92, SD = 0.78	21.49 $p < 0.001$	0.42
Blame motivation (company)	0.84	M*intentional* = 4.57, SD = 0.20 M*accidental* = 3.98, SD = 0.22	3.82 $p < 0.05$	0.35
Severity of punishment (company)	–	M*intentional* = 52.03, SD = 0.54 M*accidental* = 40.58, SD = 0.71	5.60 $p < 0.05$	0.37
Blame motivation (developer team)	0.77	M*intentional* = 5.46, SD = 0.14 M*accidental* = 4.56, SD = 0.15	12.54 $p < 0.001$	0.60
Severity of punishment (developer team)		M*intentional* = 65.90, SD = 0.36 M*accidental* = 49.49, SD = 0.58	10.04 $p < 0.01$	0.57

Standard deviation scores reported here are after the transformation

condition rated the action to be more wrong ($M = 5.13$; $SD = 1.41$) than did participants in the accidental condition ($M = 4.04$; $SD = 1.37$); F (1, 148) = 19.70; $p < 0.001$; Cohen's $d = 0.63$). There was no significant difference in lost experienced by the company (intentional condition: $M = 5.57$, $SD = 1.05$; accidental condition: $M = 5.34$, $SD = 0.93$; F (1, 148) = 1.92; $p = 0.17$; Cohen's $d = 0.12$).

We log-transformed the blame judgments scores and used them as the outcome variables in our final analyses. The correlations among blame judgment scores across entities are presented in Table 10. The significant correlations between moral judgments toward the company and AI, and between moral judgments toward developers and AI might suggest that there is some connection between blaming the creators (i.e., the company and developers) and AI—as people blame AI more for its intentionality, they also put more blame on both the company and developers.

As summarized in Table 11, people believed the AI system in the intentional condition deserved more blame than in the accidental condition. Relative to those in the accidental condition, participants in the intentional condition also assigned more punishment to the AI system. Consistent with our hypotheses, people believed both the company and developers in the intentional condition deserved more blame than those in the accidental condition.

We next compared the blame judgments scores of one entity with another, we found the blame judgments followed the same pattered as those in Experiments 1 and 2. In both conditions, people assigned more blame to developers than to either the company or the AI system (see Figs. 12 and 13). Both blame motivation and severity of punishment scores were higher for the developers than for the company and the AI system, and the differences were significant (see Table 12).

The Cronbach's Alpha (α) for perceived agency and experience were 0.72 and 0.88, respectively. The correlation between agency and experience perceptions was significant ($r = 0.24$; $p < 0.01$). There were no significant

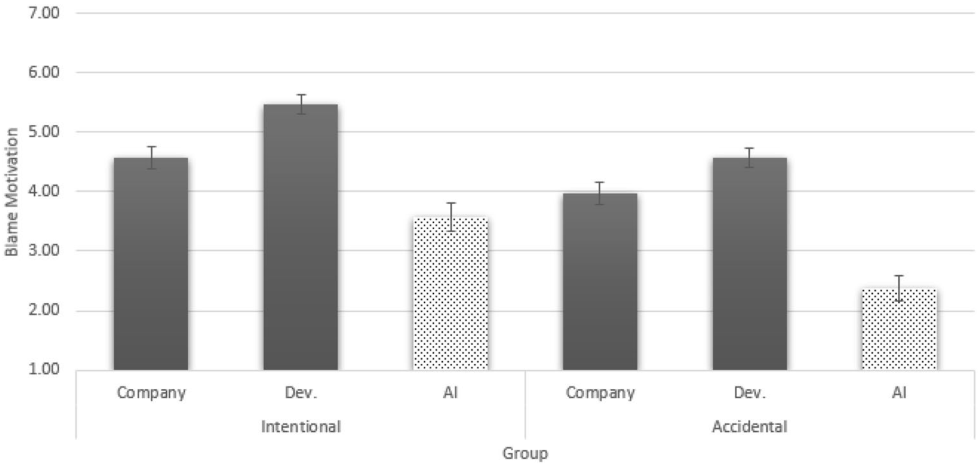

Fig. 12 Blame motivation differences among entities (Experiment 3)

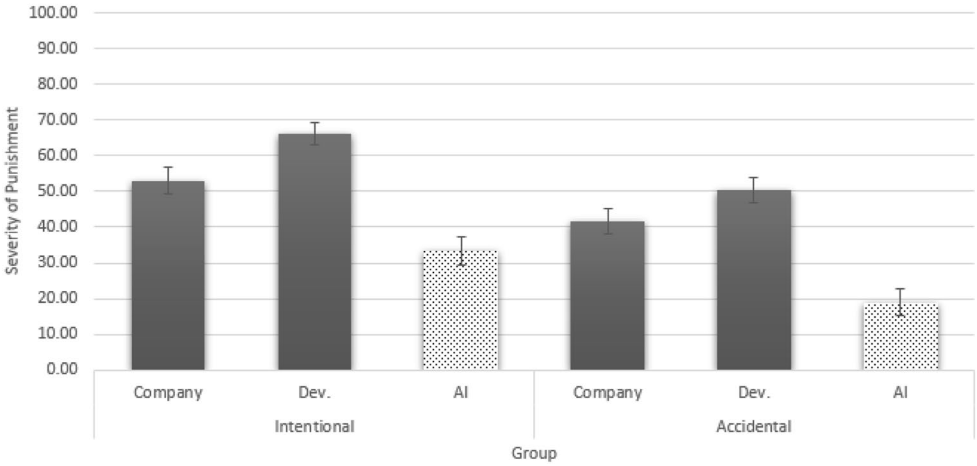

Fig. 13 Severity of punishment differences among entities (Experiment 3)

Table 12 Test of mean differences in blame judgments (Experiment 2)

Condition	Group comparison	ΔMean	t-value	p-value
Intentional	Blame movation (company vs. developer team)	− 0.10	− 5.19	< 0.001
	Blame movation (company vs. AI)	0.16	3.78	< 0.01
	Blame movation (developer team vs. AI)	0.25	6.80	< 0.001
Intentional	Severity of punishment (company vs. developer team)	− 0.20	− 4.08	< 0.001
	Severity of punishment (company vs. AI)	0.43	4.51	< 0.001
	Severity of punishment (developer team vs. AI)	0.63	7.09	< 0.001
Accidental	Blame movation (company vs. developer team)	− 0.08	− 3.61	< 0.01
	Blame movation (company vs. AI)	0.29	9.11	< 0.001
	Blame movation (developer team vs. AI)	0.37	11.12	< 0.001
Accidental	Severity of punishment (company vs. developer team)	− 0.20	− 3.47	< 0.001
	Severity of punishment (company vs. AI)	0.76	8.20	< 0.001
	Severity of punishment (developer team vs. AI)	0.95	10.63	< 0.001

Table 13 Summary of hypotheses results

Hypothesis	Experiment 1	Experiment 2	Experiment 3
H1: People will attribute higher blame judgments toward AI when a violation is perceived to be intentional than when it is perceived to be accidental	Supported	Supported	Supported
H2a: Perceived agency in AI mediates the relationship between perceived intentional harm (directed to humans) and blame judgments toward AI	Not supported	Not supported	Supported (Null Hypothesis)
H2b: Perceived experience in AI mediates the relationship between perceived intentional harm (directed to humans) and blame judgments toward AI	Supported	Supported	Supported (Null Hypothesis)
H3: People will attribute higher blame judgments toward organizations when a violation involving AI is perceived to be intentional than when it is perceived to be accidental	Not supported	Supported	Supported
H4: People will attribute higher blame judgments on developers when a violation involving AI is perceived to be intentional than when it is perceived to be accidental	Not supported	Supported	Supported

differences in every index of mind attribution across conditions: agency (intentional condition: $M=5.75$, $SD=1.02$; accidental condition: $M=5.42$, $SD=1.21$; F (1, 148) = 2.86; $p=0.09$; Cohen's $d=0.29$) and experience (intentional condition: $M=2.36$; SD = 1.06; accidental condition: $M=2.11$; $SD=1.08$; F (1, 148) = 2.23; $p=$ 0.14; Cohen's $d=0.22$).

Given these insignificant findings, we did not perform a mediation test.

Summary of Hypotheses Testing

In all three studies, we found most of our hypotheses were supported (see Table 13). H1 was supported in all experiments. However, we did not find support for H2a in all three studies, whereas we found support for H2b in Experiments 1 and 2. In Experiment 3, we found perceived intent to cause harm to a non-human entity did not increase perceptions of AI's mind. Thus, they did not mediate the relationship between perceived intent to cause harm and blame judgments toward AI. Regarding blame judgments toward non-AI entities, we found no significant differences in the blame judgment indexes attributed to the company and developer team in both conditions of Experiment 1. However, in Experiments 2 and 3, when the harm did not cause a person death, the blame judgment scores in the intentional condition were higher than in the accidental condition for both the company and developer team.

Further, Bigman et al. (2019) speculated that ascribing more mind attributions to AI will lead to less blame on its human creators and their owners. We found the relationships between perceived mind (perceived agency and perceived experience on AI) and blame judgments toward the company and developers ranged from nonsignificant to negative significant (the highest *correlation score* was the correlation between perceived experience and blame judgments toward the company and developers in Experiment 1 ($r = -0.16$; $p < 0.01$). These findings suggest that perceived mind in AI does not influence blame judgments toward other entities. Instead, blame can be distributed across multiple agents—assigning more blame to one agent doesn't guarantee less blame to another agent.

Discussion

The main goal of our research is to investigate the attributions of blame judgments assigned to multiple entities—organizations, developers, and an AI system itself when AI is involved in moral violations. In a series of experiments, we explore the following question: "who will be held responsible for harm involving an AI system?" Our studies draw upon the theory of mind perception (Gray et al., 2007) and moral judgments literature. Our findings reveal that perceived intentional harm leads to perceived mind in AI, and perceived experience in AI, but not perceived agency, mediates the relationship between perceived intentional harm caused by AI and blame judgments toward AI. We also found that when an incident causes a victim's death, people blame the company and developer team in the intentional harm condition as much as they did in the accidental condition (Experiment 1). However, when an incident only causes a minor injury (Experiment 2) or productivity lost to a company (Experiment 3), people saw perceived intentional harm as worse than accidental harm and consequently, they put more blame on the company and developer team in the intentional harm condition than in the accidental condition. We also find people blame developers the most in all scenarios, followed by companies and AI. The findings have implications for several research disciplines, including studying Human-AI interactions, designing intelligent agents, and anticipating potential unintended moral consequences of AI systems. Below, we further elaborate on the theoretical and practical implications.

Theoretical Implications and Future Research Directions

Our research has several theoretical implications. First, our research shows an AI outcome is the result of distributed agency (Taddeo & Floridi, 2018). Thus, it is possible to hold developers and companies responsible for moral violations. In all three experiments, AI is seen as a target of moral judgments, and when moral violations are perceived as causing human suffering, people impose moral judgments to AI by attributing human mind to the system. An AI system is in part of the product of an engineer and in part a self-taught machine. Although such a system does follow instructions, these instructions also tell them to be independent, learn from experience, try out new strategies, and learn from these trials' outcomes (Gless et al., 2016). Since AI draws its own conclusions, the outcomes of its action cannot be predicted in advance. An AI system might be a risk if it is designed to do something beneficial, but it takes unpredicted paths to accomplish its goals or if it is given more autonomy to make a critical decision. The mysterious mind of AI systems points to the dark side of AI. Algorithms underlying AI systems have been proven to be powerful at solving problems, and they have been widely deployed for tasks, such as image captioning, voice recognition, and information search. However, without knowing who will be responsible for the automatic actions performed and on what ground we judge harmful behaviors of an AI system, the benefits of AI are at stake. According to the dyadic morality perspective, perceived intentionality has a stronger effect than unintentionally, even if the intentional act can be explained with reasons (Gray & Wegner, 2012). We do not know yet whether blame judgments may be influenced by the nature of reasons given for an action. According to the blame judgments literature (Malle et al., 2014), blame judgments incorporate the notion of justification (Malle, 2021). Future research should investigate whether different reasons (e.g., selfish, altruistic, etc.) may influence mind attributions or blame judgments.

Second, contrary to the dyadic morality perspective, which suggests that deserving punishment for wrongdoing correlates more with perceived agency, we find that perceived experience, not perceived agency, is what define an AI system as a moral agent. We explain these findings from several perspectives. First, the findings are consistent with the most recent development of the theory of mind perception which argues that experience is necessary for moral judgments (Bigman & Gray, 2018; see also Himma, 2009). Capacity for empathy (i.e., feeling pain on behalf of others) seems to be a core element of moral judgments in the AI context. Second, AI morality could come from emotions that provide an immediate feeling of right or wrong, and moral judgments are the product of quick and automatic intuitions that then give rise to slow, conscious moral reasoning (Haidt & Bjorklund, 2008). If moral judgments are made intuitively, then evaluators might rely more on the target's mind attributions, including an affective valance (e.g., good versus bad personality), without any conscious awareness of having gone through steps of search, or inferring a conclusion from the target's agentic mind attributions. Our findings on the relationship between perceived experience and blame judgments toward AI suggest that people use different moral norms when they interact with AI.

Given morality is a social practice (Hage, 2017; Malle et al., 2019), AI systems that interact with humans are seen as deeply embedded in a social structure and treated as "human counterparts." Such systems are expected to engage in pro-social behaviors (Eisenberg & Miller, 1987). As people see an artificial agent as a social agent with unpredictable outcomes, they ascribe more attributions of emotions to that agent and expect it to have empathy—the ability to comprehend other entities' affective or cognitive status (Eisenberg & Miller, 1987). Thus, when AI is perceived to cause harm intentionally, it increases people's perceptions of AI's experience, and in turn, influence its blameworthiness.

The insignificant mediation effect of perceived agency could be because people morally evaluate humans and machines differently. Malle et al., (2019) also pointed toward this possibility. Blame is unique in many aspects, from its focus on the agent (e.g., mental state inferences) to its broad range of information processing (e.g., norms, causality, intentionality, and reasons). Different processes of blame judgments between humans and AI may arise from different perceptions of their social roles and moral justifications that come with those roles (Malle et al., 2019). When people interact with AI, they might view the system as a social agent in different ways, expecting the system to engage in moral behaviors as if they have human nature or experience (Bastian et al., 2011). Future research is needed to explore different theoretical explanations of why experience mediates the relation between perceived intentional harm and blame judgments toward AI, whereas perception of agency does not.

Third, we performed an analysis of an AI's life-cycle to identify potential responsible parties and suggested that other parties, including the company that owned the system and developers who designed it, can be held responsible when AI is perceived to be the cause of a moral violation. Although participants' views about how an AI system can be punished might vary as they might use different justifications when they judge an AI system, there is sufficient evidence from prior research to suggest that AI may generally be attributed less fault than humans. For example, Voiklis et al. (2016) found that AI was blamed less and attributed less wrongness than humans when it made the same moral choices in a trolley-problem ethical dilemma. In contrast, AI can sometimes be blamed more than humans [e.g., when AI teammates make mistakes (Merritt et al., 2011) (see also Shank et al., 2019)]. AI is usually blamed more than humans when AI is expected to act in a more efficient, optimal, and rational way than humans can do (Shank et al., 2019). In our research, we found that AI was blamed less than both the company and developers. Without a legal framework to deal with an AI system's liability, a victim can easily place the blame on the nearest responsible parties involved in an AI life-cycle (Hao, 2019b). Regulators should adopt standards that would help distribute responsibility fairly. For example, it could be accomplished by developing standards specifying the characteristics of AI systems should have, such as being limited to specified activities (Scherer, 2016).

Fourth, we found that mind attributions seem to be induced by perceived intentional harm targeting a human entity, but not to a non-human (or possibly non-living) entity (e.g., a company's productivity lost). This could be because perceived harm to a person is expected to have more severe consequences than harm to a non-human entity. As Waytz et al. (2010) suggested, AI systems that cause negative outcomes seem to be attributed with higher mind quality than systems that produce positive outcomes. These findings suggest that people often make an intuitive judgment when the harm caused by AI threatens individuals' lives, that agent might have a mind like a human. To the extent that the threats do not cause human suffering, people seem to perceive an AI system as merely a smart machine with a certain degree of agency. When people sense the threats may harm other people, an artificial agent is positioned as a moral subject.

Fifth, we demonstrate the relevance of studying the blameworthiness of AI from the attributional perspective. Although AI does not have real intentionality like humans, people do perceive AI systems as having a human mind when they observe these systems behave or act like a human agent (Hage, 2017). Across all studies, we found this following statement to be true—attributing mind to an AI system seems to be induced by perceived intentional harm. It could be because perceived intentional harm is tied to negative consequences resulting from the system's unpredicted behaviors.

Lastly, we respond to the calls for the need to understand the consequences of AI. One main feature of AI systems is that they can act autonomously without human interventions. Many AI algorithms represent "black boxes" when it comes to understanding how results are produced (Seeber et al., 2020). As technologies advance, it is critical for researchers to discover how we can work on explainable AI and understand who will be counted accountable if harm is perceived.

Limitations and Future Research Directions

Our studies are not without limitations. First, we only asked our participants to examine the dimensions of mind of AI. Future research should also include other agents (e.g., human adults, animals, etc.) as Gray et al. (2007) did in their study. It will be valuable to include mind perceptions of various agents (e.g., corporations, developers) and compare the scores of their mind perceptions. Second, we only used samples from the United States. The findings from our studies can be further expanded using a broader base of the population across cultures. Third, one type of autonomous agent was chosen and was experimentally manipulated for its actions in words. The disadvantage of this procedure is observers can construct their own physical image of the artificial agent. Additional research could examine the impact of different types of autonomous agents and their physical appearances on mind perceptions and moral judgments. Fourth, our experiments were conducted at a specific point in time. While we were able to manipulate perceived intentional harm and demonstrate its correlation with mind perceptions, a reversed direction is possible. Future research using a longitudinal study design is needed to strengthen our theoretical model. Fifth, it is possible that assigning a gender to an AI system (e.g., AI is assigned a male name) influenced participants' judgments on the moral violation of an AI system. Although recent research (e.g., Capraro & Sippel, 2017) suggests that moral dilemmas are driven by emotional salience, not by gender differences in moral violations, future research should control for the effect of assigning a gender to AI on moral judgments. Lastly, we used the same scale to measure blame judgments toward various parties. Although the correlations between the blameworthiness rates of AI and non-AI entities ranged from small to not significant, we acknowledge that there is a possibility that participants made comparative judgments about who deserved more punishment. Problems might arise if people can more readily imagine the nature of punishment given to a human as opposed to an AI system. Future research might consider using a different scale to measure blame judgments of AI and non-AI entities.

Practical Implications

Investigating the relationship between humans and AI systems from a moral perspective has several practical implications. First, our studies demonstrate how blameworthiness is attributed to multiple entities when a moral violation is perceived to be committed by an AI system. The findings show that people do hold parties with power (i.e., companies) and the creators (i.e., developers) responsible for a moral violation involving AI. In an extreme case, as shown in Experiment 1, people blame these parties equally both in the accidental condition and in the intentional condition. Although there is a concern that as people ascribe more mind attributions to AI, they will put less blame on their human creators or their owners (Bigman et al., 2019), our studies show that this is not the case. The motivation to blame corporations and developers is high across all studies, regardless of whether perceived harm is intentional or accidental. These findings suggest that people see corporations as legal and economic entities constructed to pursue social and economic objectives (Ashman & Winstanley, 2007) and respect human dignity above everything else (Brusoni & Vaccaro, 2017). The motivation to blame and punish AI, however, is new. Firms and policymakers should create a cross-disciplinary research environment that encourages researchers to investigate AI accountability and liability further.

Second, our research shows how mind perceptions and attributional processes influence how people evaluate an AI system's behaviors. As people believe an AI system possesses some characteristics of a human's experience mind, they are likely to attribute moral responsibility to the system. Although the theory suggests that ascribing the agentic mind to an agent may qualify that agent as a moral agent, our findings do not find support for this claim. Instead, our studies suggest that attributing agency mind to AI doesn't contribute to its moral judgments, even when the actions are perceived to be intentional. Based on these findings, we recommend organizations and the creators of AI to focus on promoting AI's agentic characteristics. Depending on the context in which AI is used, organizations may label an AI product or service its agentic features and limitations and not misrepresent an AI system as a human. The purpose of designing AI is to ensure AI is safe, secure, and susceptible to human control.

Lastly, our study provides initial evidence to support the consensus that people do hold AI systems responsible if they believe the harm is intentional and targeted to a person. By understanding how human minds make sense of morality and how people perceived mind of an AI system, our findings help AI designers and engineers understand how their creations are likely to be perceived. As they put more autonomy on an AI system, they should carefully consider the implication of their design decision on morality. To the extent that scientists and policymakers are concerned with public opination, they may have to be prepared to face ethical and legal issues that humanity has never faced before. Based on our research findings, we argue that our immediate goal is to design and create AI systems that are more sensitive to ethically important aspects of their tasks. For example, although AI may or may not be capable of understanding right or wrong, it would be extremely valuable for the system to understand and share the feelings of others so they can perform in a morally acceptable manner. Since people tend to rely on the experience mind when they morally judge an artificial agent, developing a broad array of moral considerations into an agent's choices and actions needs to be our priority.

Conclusions

To conclude, our studies demonstrate how blame is attributed to multiple parties when a moral violation involving AI takes place. Our findings demonstrate the impact of perceived intentional harm on the ascription of mental states to an AI system and how these mind perceptions (especially perceived experience) mediate the relationship between perceived intentional harm and blame judgments. Noteworthy, this research focuses on autonomous elements of an AI system from the social and psychological perspectives, rather than on more commonly investigated humanlike features such as appearances and voices. Moral and ethical values add interest and complexity; they are inevitable components of research that the business field should deal with as the adoption of AI increases. We hope this research can provide an impetus for other studies in the realm of designing an accountable AI system, building on theoretical and practical contributions as presented here.

Declarations

Conflict of interest We have no conflicts of interest to disclose.

Ethical Approval The three studies were granted exemption by the first author's home institution according to federal regulation 45 CFR 46.104(d)(2): Research involving the use of educational tests, survey procedures, interview procedures, or observation of public behavior. We certify that all procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the studies.

References

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., & Pizzato, L. (2020). Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, *30*(1), 127–158.

Adams, A., & Sasse, M. A. (1999). Users are not the enemy. *Communications of the ACM, 42*(12), 40–46.

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24*(9), 1755–1762.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183–189.

Arkin, R. C., & Ulam, P. (2009). An ethical adaptor: Behavioral modification derived from moral emotions. *2009 IEEE international symposium on computational intelligence in robotics and automation-(CIRA)* (pp. 381–387). IEEE.

Ashman, I., & Winstanley, D. (2007). For or against corporate identity? Personification and the problem of moral agency. *Journal of Business Ethics, 76*, 83–95.

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology, 50*(3), 469–483.

Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. M. (2012). Don't mind meat? The denial of mind to animals used for human consumption. *Personality and Social Psychology Bulletin, 38*(2), 247–256.

BBC. (2014). *Stephen Hawking warns artificial intelligence could end mankind*. Retrieved from https://www.bbc.com/news/technology-30290540

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines, 30*, 195–218.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

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.

Brusoni, S., & Vaccaro, A. (2017). Ethics, technology and organization innovation. *Journal of Business Ethics, 143*, 223–226.

Capraro, V., & Sippel, J. (2017). Gender differences in moral judgment and the evaluation of gender-specified moral agents. *Cognitive Processing, 18*(4), 399–405.

Cohen, S. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The Claremont symposium on applied social psychology* (pp. 31–67). Sage.

Coombs, C., Hislop, D., Taneva, S. K., & Barnard, S. (2020). The strategic impacts of Intelligent Automation for knowledge and service work: An interdisciplinary review. *The Journal of Strategic Information Systems, 29*(4), 101600.

Courthousenews. (2017). Case Case 1:17-cv-00219 ECF No. 1 filed 03/07/17, Available online: https://www.courthousenews.com/wp-content/uploads/2017/03/RobotDeath.pdf.

Donald, S. J. (2019). Don't blame the AI, it's the humans who are biased. *Toward Data Science*. Retrieved at https://towardsdat ascience.com/dont-blame-the-ai-it-s-the-humans-who-are-biased-d01a3b876d58

Doyle, C. M., & Gray, K. (2020). How people perceive the minds of the dead: The importance of consciousness at the moment of death. *Cognition, 202*, 104308.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin, 101*(1), 91–119.

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Gless, S., Silverman, E., & Weigend, T. (2016). If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Criminal Law Review, 19*(3), 412–436.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046.

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

Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind perception in psychopathology. *Proceedings of the National Academy of Sciences of the United States of America, 108*(2), 477–479.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505–520.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*, 125–130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*(2–3), 115–121.

Hage, J. (2017). Theoretical foundations for the responsibility of autonomous agents. *Artificial Intelligence and Law, 25*(3), 255–271.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814.

Haidt, J., & Bjorklund, F. (2008). Social intuitionists answer six questions about moral psychology. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 181–217). Cambridge, MA: MIT Press.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*(4), 613–628.

Hao, K. (2019a). This is how AI bias really happens—And why it's so hard to fix. *MIT Technology Review*.

Hao, K. (2019b). When algorithms mess up, the nearest human gets the blame. *MIT Technology Review*.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29.

Hollebeek, L. D., Sprott, D. E., & Brady, M. K. (2021). Rise of the machines? Customer engagement in automated service interactions. *Journal of Service Research, 24*(1), 3–8.

Hume, D. (1751). *An enquiry concerning the principles of morals*. Clarendon Press.

Ishizaki, K. (November, 2020). *AI model lifecycle management: Overview*. IBM, Retrieved from https://www.ibm.com/cloud/blog/ai-model-lifecycle-management-overview

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.

Johnson, D. G., & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology, 20*(4), 291–301.

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., & Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction* (pp. 33–40). IEEE.

Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition, 146*(1), 33–47.

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences, 7*(1), 67–83.

Kozhaya, J. (November, 2020). *AI model lifecycle management: Build phase*. IBM, Retrieved from https://www.ibm.com/cloud/blog/ai-model-lifecycle-management-build-phase

KPMG. (2020). *Avoiding setbacks in the intelligent automation race*. Retrieved from https://advisory.kpmg.us/content/advisory/en/index/articles/2018/new-study-findings-read-ready-set-fail.html

Lagioia, F., & Sartor, G. (2020). Ai systems under criminal law: A legal analysis and a regulatory perspective. *Philosophy & Technology, 33*, 433–465.

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents? The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human–computer Studies, 64*(10), 962–973.

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research, 120*, 262–273.

Malle, B. F. (2019). How many dimensions of mind perception really are there? *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 2268–2274). Cognitive Science Society.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology, 72*, 293–318.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*, 147–186.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Robotics and well-being* (pp. 111–133). Springer.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. in *2014 IEEE international symposium on ethics in science, technology and engineering, ETHICS.* IEEE.

Merritt, T. R., Tan, K. B., Ong, C., Thomas, A., Chuah, T. L., & McGee, K. (2011, March). Are artificial team-mates scapegoats in computer games. In: *Proceedings of the ACM 2011 conference on computer supported cooperative work*, pp. 685–688.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*(2), 215.

Mou, X. (2019). Artificial Intelligence: Investment trends and selected industry uses. *IFC EMCompass Emerging Markets, 71*, 1–8.

Omohundro, S. M. (2008). The basic AI drives. In: *Artificial General Intelligence*, pp. 483–492.

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society, 23*(5), 719–735.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891.

Pyszczynski, T., Greenberg, J., & Solomon, S. (1997). Why do we need what we need? A terror management perspective on the roots of human social motivation. *Psychological Inquiry, 8*(1), 1–20.

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-generation digital platforms: Toward human–AI hybrids. *MIS Quarterly, 43*(1), iii–ix.

Rai, T. S., & Diermeier, D. (2015). Corporations are cyborgs: Organizations elicit anger but not sympathy when they can think but cannot feel. *Organizational Behavior and Human Decision Processes, 126*, 18–26.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Rybalko, D. (November, 2020). *AI model lifecycle management: Deploy phase*. IBM, Retrived from https://www.ibm.com/cloud/blog/ai-model-lifecycle-management-deploy-phase.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law & Technology., 29*(2), 354–400.

Schraube, E. (2009). Technology as materialized action and its ambivalences. *Theory & Psychology, 19*(2), 296–312.

Seeber, I., Waizenegger, L., Seidel, S., Morana, S., Benbasat, I., & Lowry, P. B. (2020). Collaborating with technology-based autonomous agents: Issues and research opportunities. *Internet Research, 30*, 1–18.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society, 22*(5), 648–663.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.

Tang, S., & Gray, K. (2018). CEOs imbue organizations with feelings, increasing punishment satisfaction and apology effectiveness. *Journal of Experimental Social Psychology, 79*, 115–125.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & Society, 22*(4), 495–521.

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

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. robot agents. *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 775–780). IEEE.

Wallach, W., Allen, C., & Franklin, S. (2011). Consciousness and ethics: Artificially conscious moral agents. *International Journal of Machine Consciousness, 3*(01), 177–192.

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science, 24*(8), 1437–1445.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*, 383–388.

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.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: The MIT Press.

Wegner, D. M., & Gray, K. (2017). *The mind club: Who thinks, what feels, and why it matters*. Penguin Random House.

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2020). Robots at work: People prefer—And forgive—Service robots with perceived feelings. *Journal of Applied Psychology, 106*(10), 1557–1572.