AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

In the public domain
ISSN: 2325-9965

2025, Vol. 12, No. 4, 287–314
https://doi.org/10.1037/dec0000264

Comparing Ethical Decision Making Between Human and Machine Referents

Gene M. Alarcon[1], August Capiola[1], Krista N. Harris[2], Scott K. Meyers[2],
Sarah A. Jessup[1], and Jacob Noblick[3]

[1] Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio, United States
[2] DCS Corporation, Dayton, Ohio, United States
[3] GDIT, Dayton, Ohio, United States

Recent research and technological advances have raised questions pertaining to ethics in human–machine interaction, particularly users' evaluation of machine decision-makers and how this may differ from their evaluation of human decision-makers. The literature is notably sparse on actual experiments on how people view ethical decision making from a machine compared with a human, particularly when contextual features of a given scenario provide contradicting assumptions—from a utilitarian versus deontological perspective—on what *ought* to be done. The current research utilized the consequences, norms, and inaction model to explore differences in participants' perceptions of human or machine decision-makers with 12 new scenarios when utilitarian and deontological norms toward (in)action are (not) aligned. Across two studies, we demonstrated participants' preference for machines making utilitarian ethical decisions over humans. However, this was qualified by a participant's preference for action; that is, in Study 1, those with a tendency for *inaction* were *more* likely to endorse an action if the referent was a machine and the benefits of a decision were greater than its costs. Participants also were more accepting of normative violations from a human than a machine in Study 1. Results from Study 2 largely replicate Study 1 but also demonstrate ethical decision making may not be binary as the consequences, norms, and inaction model suggests and may require continuous criteria to capture this nuance in decision-making processes when individuals are paired with human and machine referents.

Keywords: ethical decision making, bias, robots, action–inaction asymmetry, multinomial processing tree

Supplemental materials: https://doi.org/10.1037/dec0000264.supp

Tim Rakow served as action editor.

Gene M. Alarcon https://orcid.org/0000-0003-4093-8792

The views expressed are those of the authors and do not reflect the official guidance or position of the U.S. Government, the Department of Defense, or of the U.S. Air Force, Distribution A, approved for public release: AFRL-2023-4532, cleared on Sep 14, 2023.

Gene M. Alarcon played a lead role in conceptualization, data curation, investigation, methodology, project administration, supervision, and writing–original draft. August Capiola played a supporting role in conceptualization, data curation, and writing–original draft. Krista N. Harris played a supporting role in conceptualization, data curation, and writing–original draft. Scott K. Meyers played a supporting role in conceptualization, data curation, and writing–original draft. Sarah A. Jessup played a supporting role in conceptualization, data curation, and writing–original draft. Jacob Noblick played a supporting role in data curation.

Correspondence concerning this article should be addressed to Gene M. Alarcon, Air Force Research Laboratory, 2210 Eighth Street, Building 146, Wright-Patterson Air Force Base, OH 45433, United States. Email: gene.alarcon.1@us.af.mil

Complex machine systems are increasing in capability. With the popularized integration of artificial intelligence and machine learning into a litany of systems, usage has increased in a variety of contexts (e.g., large language models for web browsing, Roumeliotis & Tselikas, 2023; targeted advertising, Kietzmann et al., 2018; online video suggestions, Liao et al., 2021). Currently, machines are being leveraged in areas such as parole decisions, logistics, health care, and manufacturing (L. Huang & Peissl, 2023; Wu et al., 2020). The growing application of machines to a variety of contexts has led to increased discussion over the philosophical and practical implications of machines making ethical decisions (Pflanzer et al., 2023).

With unmanned machines and other autonomous systems entering the workforce, machines will face ethical dilemmas as humans do in these contexts (Pflanzer et al., 2023). The rapid advancements of machines in a variety of scenarios requiring ethical decisions necessitate an understanding of how humans perceive the ethical decision making of machines. Although the application of machines to ethical decision making may seem far off, there are current instances, such as self-driving cars (Awad et al., 2018), and deployment of large language models (Stahl & Eke, 2024), such as ChatGPT, that illustrate machines make decisions that have ethical implications.

Researchers and developers could simply apply the same ethical expectations to machines that are expected of humans, but it remains to be seen if the ethical decisions of a human action and machine action are viewed as identical when perceived by a human. The current research explored differences in peoples' perceptions of machines and humans making ethical decisions. We utilized and expanded on previous methods (Gawronski et al., 2017) to differentiate between utilitarian and deontological norms. Across two studies, we explored whether participants perceived ethical decision making by humans and machines differently depending on consequences and normative constraints of a scenario.

Normative Ethics Theories

Normative ethics is a branch of philosophy focused on human conduct, specifically the behaviors of agents in society. Although the field of normative ethics is quite diverse, popular discourse and debate are largely demarcated into two areas: consequentialism and deontology (Gawronski et al., 2017). Consequentialism is a theory of normative ethics that advocates behaviors should be rated on the outcomes of the behavior (Scheffler, 1988). Arguably, the most popular form of consequentialism is utilitarianism, which advocates an action is ethical if the action results in the best possible benefit for the most people (Mill, 1863). In contrast, an action is unethical if it results in comparatively less benefit overall.

Deontology is the normative ethical theory that posits an action is ethical based on the action itself, which is outlined by a series of duties and obligations (Waller, 2005). There are several different deontological ethical theories such as divine command theory (Hare, 2015) and Ross' deontological pluralism (Ross, 2002), but the most well-known deontological theory is Kantianism (Kant, 1785/1873). Kantianism states that certain actions are intrinsically good or bad. A main theorem of Kantian ethics is the categorical imperative which has two maxims: act only in accordance with what you would want to become universal law and never treat humanity as a means to an end but rather always as the end. For example, Kantian ethics state that lying is an immoral action, even if the lie is to protect someone's life because one would not advocate lying as a universal maxim. Although we acknowledge there are several other deontological theories, we use the term deontology to refer to Kantian deontology onward in this article for clarity.

Measuring Ethics

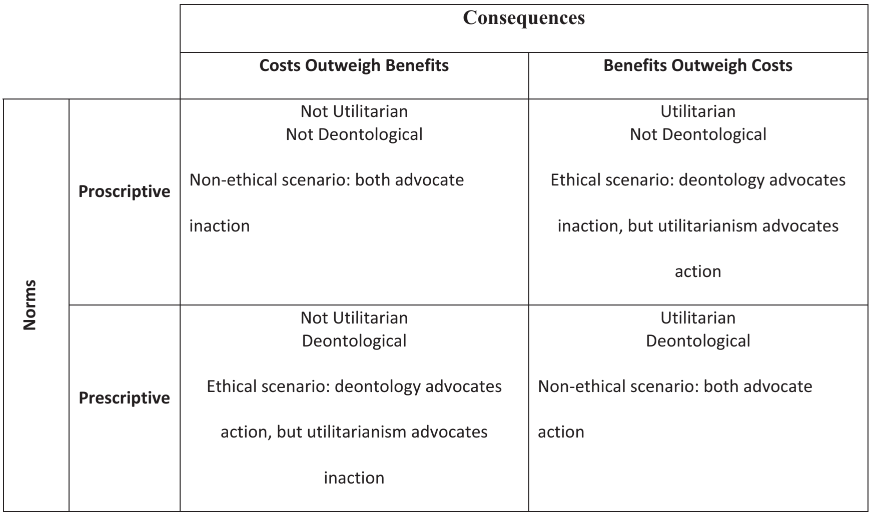
Utilitarianism and deontology are concerned with the ethical actions a moral agent performs. While the former is concerned with rules that dictate how to act given the consequences of the action, the latter is concerned with the actions themselves (Tseng & Wang, 2021). Philosophers often create thought experiments to test the philosophical rigor of ethical theories. One popular thought experiment is the trolley problem (Foot, 1967). The trolley problem posits a scenario in which a runaway trolley is on course to collide with several people (usually five) some distance down the track. The ethical agent can divert the trolley onto another track, but this would kill one person on the other track. Essentially, the scenario is used to explore the options of doing nothing and letting a certain number of people die compared with intervening and sacrificing one person to save several lives. This philosophical dilemma has been used repeatedly in the literature to explore moral perceptions of human agents (Bauman et al., 2014) and autonomous agents (Yokoi & Nakayachi, 2021).

Although the trolley problem is relevant to certain scenarios such as autonomous vehicles, it does have drawbacks. First, it has been argued the scenario is too extreme and disconnected from real life to be useful (Khazan, 2014). Although the trolley problem has been utilized in investigating preferences for the programming of autonomous vehicles (Yokoi & Nakayachi, 2021), it is not an accurate representation of most ethical dilemmas and may be too specific to apply to other ethically charged human–machine interactions (see our scenarios in Supplemental Materials). Second, Gawronski et al. (2017) noted in the standard illustration of the trolley problem, decisions aligning with deontological norms require inaction (do not switch the tracks). In contrast, a decision aligning with a utilitarian ethic requires action (switch the tracks). Gawronski et al. mention that in these scenarios, decisions aligning with a deontological ethical perspective cannot be separated from a response tendency toward inaction, whereas decisions aligning with a utilitarian ethical perspective cannot be separated from a response tendency toward action (see also Körner et al., 2020).

To address some of the concerns of conflating action with utilitarianism and inaction with deontology, Gawronski et al. (2017) used multinomial processing trees (MPTs) to demarcate sensitivity to consequences, sensitivity to norms, and the general preference for inaction regardless of consequences or norms, which they called the consequences, norms, inaction (CNI) model. MPTs separate complex cognitive processes into categorical constructs (Singmann & Kellen, 2013); in the current case, consequences, norms, and inaction are the categorical constructs. In Gawronski et al.'s experimental paradigm, participants are given scenarios with four kinds of ethical dilemmas, which capture two orthogonal factors: consequences (the benefits of a decision are greater than its cost; the costs of a decision are greater than its benefits) and norms (prescribing action aligns with moral norms; prohibiting action aligns with moral norms). These quadrants are illustrated in Figure 1. Figure 2 illustrates the CNI model as MPT. The CNI model estimates probabilities for consequences, norms, and action/inaction preference driving endorsement in each of the four scenarios. MPTs can describe the response frequencies based on a multinomial distribution. Each latent construct is obtained

Figure 1

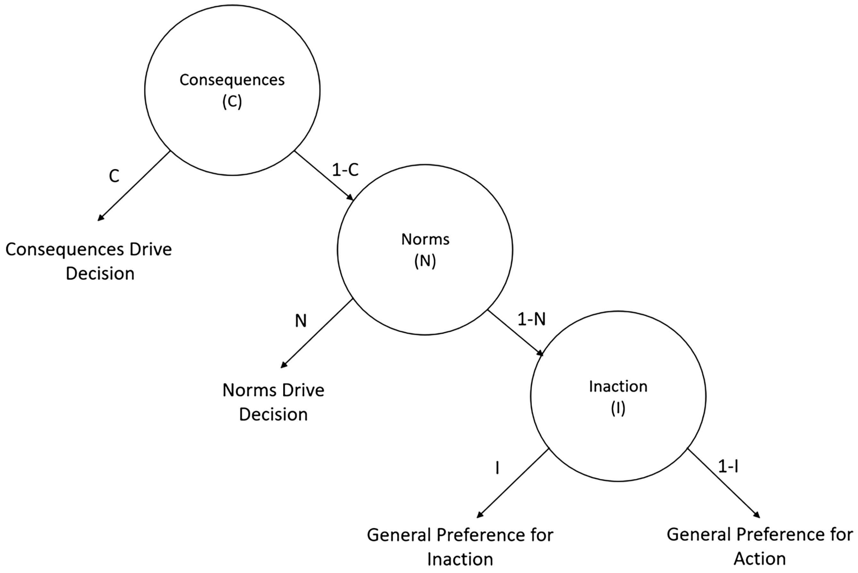
The Four Quadrants of the Ethical Scenarios as a 2 × 2 Design of Deontology and Utilitarianism



		Consequences	
		Costs Outweigh Benefits	**Benefits Outweigh Costs**
Norms	**Proscriptive**	Not Utilitarian Not Deontological Non-ethical scenario: both advocate inaction	Utilitarian Not Deontological Ethical scenario: deontology advocates inaction, but utilitarianism advocates action
	Prescriptive	Not Utilitarian Deontological Ethical scenario: deontology advocates action, but utilitarianism advocates inaction	Utilitarian Deontological Non-ethical scenario: both advocate action

Figure 2

Multinomial Processing Tree for Consequence, Norms, and Inaction Model


Consequences (C)
C
1-C
Consequences Drive Decision
Norms (N)
N
1-N
Norms Drive Decision
Inaction (I)
I
1-I
General Preference for Inaction
General Preference for Action

by combinations of cognitive processes, which are assumed to occur in an all-or-nothing fashion (i.e., Is this acceptable? Yes or no). These analyses are used to assess sensitivity to consequences (e.g., utilitarianism), sensitivity to norms (e.g., deontology), and a general preference for inaction regardless of consequences or norms. As illustrated in Figure 1, there is a probability associated with endorsing a scenario if the decision is driven by consequentialist thinking, resulting in a response probability estimated by "C" in Figure 1. If another process drives the decision, the estimates are calculated working down the MPT. For example, a normative driven response is calculated as (1 − C) × (N). Higher parameter scores on the C or N parameter indicate higher sensitivity to the underlying construct. Higher parameter scores on the I construct indicate a preference for inaction while lower scores indicate a preference for action, with scores around .50 indicating no preference for inaction or action. The CNI model has been used extensively to explore question framing (Gawronski et al., 2017), language use (Białek et al., 2019; Nadarevic et al., 2021), power (Gawronski & Brannon, 2020), and stress (Li et al., 2021; L. Zhang et al., 2018) in moral dilemmas.

Ethics and Machines

Machines enabled with artificial intelligence or machine learning are increasingly being used in ethical scenarios. As such, a dilemma arises that could result in normative ethical ramifications, for example, autonomous driving (Grigorescu et al., 2020), military applications (de Swarte et al., 2019; Stanley-Lockman, 2021), and legal decisions (Atkinson et al., 2020). Machines have been regarded as companions (Dautenhahn et al., 2005), indicating they are not being used simply for functional processes. The use of autonomous machines in a variety of applications has led to ethical questions such as "Who is responsible when artificial intelligence violates the law or ethics?" (Bartneck et al., 2021), "What are the ethical guidelines for artificial intelligence?" (Jobin et al., 2019), and "What principles should guide artificial intelligence?" (L. Huang & Peissl, 2023). As ethical guidelines are established for artificial intelligence/machine learning-enabled machines, it is important to note how actions stemming from these guidelines are interpreted by humans. Instantiating ethical guidelines that machines adhere to is a challenge in and of itself, but quantifying how these guidelines are interpreted by humans through the lens of human bias is an arduous task relevant to both philosophers, scientists, and product developers.

Humans often have higher standards for machines than for humans, leading to a “perfect” automation schema that bias perceptions of machine performance as well as the threshold for abandoning those machines should they be perceived to err (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007). Research has demonstrated people view a performance degradation from a machine more harshly than from a human, even if its performance degradation is instantiated with the same stimuli (Alarcon et al., 2023). The research on ethical violations from a machine is relatively nascent as machines have only started to become autonomous in the past decade. Researchers have compiled a list of norms for machines to operate (L. Huang & Peissl, 2023) that are very similar to norms of human behavior. For example, one of the major principles for machines is nonmaleficence (Floridi & Cowls, 2022; Motloba, 2019), which states the machine should “first, do no harm” and avoid imposing risks of harm to others. Stahl (2021) noted principles of justice and fairness are also pertinent to machines, helping to reduce bias and discrimination. Some research indicates humans apply the same moral norms to machines as they do humans (Komatsu, 2016). However, as with humans, machines will encounter ethical dilemmas when norms and consequences are at odds with each other. It remains to be seen if there are biases in how humans view the ethical behaviors of machines in comparison to ethical behaviors of humans.

The research on perceptions of ethical machine decision making is sparse and mixed. Young and Monroe (2019) found people tend to assess machines more harshly than humans. Researchers have also found people resist artificial intelligence that makes ethical decisions (M. K. Lee, 2018; Longoni et al., 2019). In contrast, others have demonstrated people had less negatively affective responses toward machines compared with humans in a modified version of the trust game where there were performance and morality aspects added to the game and the violation was morality based (e.g., the machine partner returned less money than promised to participants after the participants loaned money to the machines in the task and there were no performance issues; Alarcon et al., 2024). Alarcon et al. (2024) suggested this may have occurred because humans have different schemas for machines. These schemas may reflect an underlying belief that machines do not have agency. As a result, when an ethical norm is violated, humans may face greater consequences since they are considered moral agents. Consequently, a human breaking a normative rule may be viewed more severely than a machine committing the same violation (Alarcon et al., 2024).

An issue with most previous research is the failure to demarcate the norms from the consequences in the ethical scenarios presented to the participants. People apply norms differentially to humans and machines, such that users expect machines to make utilitarian choices instead of deontological ones (Malle et al., 2015; Voiklis et al., 2016). Throughout several iterations of a modified trolley problem, Yokoi and Nakayachi (2021) demonstrated people prefer autonomous cars that share their own utilitarian ethical view rather than a deontological ethical view. These studies support the postulate that humans have different expectations of machines, possibly stemming from their presumption that machines do not possess emotional experiences (see also H. M. Gray et al., 2007; J. D. Lee & See, 2004, p. 66). People perceive machines leveraging artificial intelligence as more likely to make utilitarian decisions than humans (Z. Zhang et al., 2022). This aligns with research indicating that people perceive machines to possess competence but little warmth, even when they are designed for social interaction (H. M. Gray et al., 2007; Liu et al., 2021). As machines are not expected to have agency (J. D. Lee & See, 2004), machines demonstrating calculated (utilitarian) rather than moral (deontological) decisions may better align with people’s schemas of machines. However, measuring specific features of an ethical scenario, while side-stepping confounds of people’s tendency for action or inaction (Gawronski et al., 2017), is a viable step in research on ethics in human–machine interaction.

Study 1

Previous research on machine ethics has focused on the permissibility of machines to perform an ethical action, with researchers noting that humans have a bias against machines making any ethical decisions (K. Gray & Wegner, 2012). However, machines are being utilized in new areas that create ethical concerns (e.g., L. Huang

& Peissl, 2023; Wu et al., 2020). The question then becomes whether humans perceive the ethical decision from a machine the same as an ethical decision from a human. Additionally, previous research has treated machine ethical decision making as a uniform process, without differentiating between utilitarian and deontological decision making (Dawtry & Callan, 2024; K. Gray & Wegner, 2012). This is problematic as researchers have found humans have a bias toward blaming machines more for not performing utilitarian behaviors (Komatsu et al., 2021; Malle et al., 2015).

To understand the possible mechanisms for the human biases toward machine consequentialism, we look to the literature on the theory of mind. H. M. Gray et al. (2007) found two dimensions through which humans perceive the minds of others. The first dimension they labeled was agency, which is the ability of the referent to think, reason, plan, and carry out intentions (K. Gray & Wegner, 2012). The second dimension they labeled was experience, which is the ability of the referent to feel sensations and emotions. Research on these dimensions has noted that humans believe machines have some degree of agency, but they lack experience (i.e., emotions; H. M. Gray et al., 2007; K. Gray & Wegner, 2012). This lack of complete experience, especially for thought, indicates machines are not able to make deontological ethical decisions (Floridi & Sanders, 2004; Hellström, 2013; Malle & Scheutz, 2014; Steinert, 2014; Wallach et al., 2010). However, researchers have noted people prefer machines to make consequentialist decisions (Komatsu et al., 2021; Malle et al., 2015).

Machines lack the requisite emotions to perform complete ethical decision making (K. Gray & Wegner, 2012). However, as machines interact with the environment, they have a degree of agency. Although people ascribe a degree of agency to machines, it is less than they ascribe to other humans (Bigman & Gray, 2018; H. M. Gray et al., 2007; K. Gray & Wegner, 2012; Yam et al., 2021). The ascription of agency to machines occurs for two reasons. First, machines can exert agency on the environment and often have anthropomorphic qualities such as human bodies, engendered voices, and displayed emotions (Epley et al., 2007). Second, people are motivated to explain and predict a machine's behavior, and ascribing the machine agency helps the person comprehend the interaction with known human mental abilities (Waytz et al., 2010). This ascription of agency but not emotions to the referent machine results in a preference for consequentialism because there is less emotional aspect to consequentialist theories (Manfrinati et al., 2013). In contrast, normative theories such as deontology focus on mores (the essential or characteristic customs and conventions of a community) rather than consequences (Greene et al., 2001). These mores have been more strongly linked to emotion in the literature than consequentialist actions (Greene, 2008; Greene et al., 2001, 2009). Researchers have noted that moral decisions are based on emotions but failed to demarcate different philosophies such as consequentialism and deontology (Bastian et al., 2012; K. Gray et al., 2012). Given the complexity of a deontological decision, which requires both agency and emotional experience, we hypothesize people will prefer a human to make a deontological decision over a machine, whereas people will prefer a machine to make a utilitarian decision over a human.

> *Hypothesis 1: Participants will perceive utilitarian actions as more ethical from a machine than a human.*
>
> *Hypothesis 2: Participants will perceive deontological actions as more ethical from a human than a machine.*

In Study 1, we explore the differences of human and machine decision making in ethical scenarios using the CNI to demarcate the preferences for consequence- and norm-driven actions. The above hypotheses focus on main effects such that people will have a preference for a machine to make utilitarian decisions and humans to make deontological decisions. The CNI model discussed in the introduction notes four conditions in the experimental paradigm. Figure 1 illustrates the four conditions of the ethical scenarios when factored by deontology and utilitarianism. In the upper left corner, the cell is characterized by both a prohibitive response from deontology and utilitarianism. As neither ethical theory would advocate action in that cell, the condition represents the least ethical decision in the four conditions. Conversely, in the lower right condition, the cell is characterized by action from both deontology and utilitarianism. In other words, both ethical theories advocate action and thus the most ethical decision of the four conditions. The complexity lies in the other two conditions where there is an ethical dilemma between the two theories, in the lower left and upper right quadrants. As such, we hypothesize there will be no differences between the machine or human referent in the quadrants where there is no ethical dilemma. Consequentialism will motivate the response for the machine referent and deontology will motivate the response for the human referent. We expect to see differences between humans and machines in the conditions where there is an ethical dilemma between utilitarianism and deontology. In these instances, we theorize utilitarianism will drive the motivation for the machine, in that people will expect the machine to follow the consequentialist theory. In contrast, with a human referent we theorize deontology will drive the motivation for action for humans, in that people will expect the human to follow deontological theory.

> *Hypothesis 3: When there is an ethical dilemma between utilitarianism and deontology, participants will perceive utilitarian actions as more ethical from a machine than a human.*

> *Hypothesis 4: When there is an ethical dilemma between utilitarianism and deontology, participants will perceive deontological actions as more ethical from a human than a machine.*

Study 1: Method

Participants

Participants (N = 543) were recruited online via Amazon's Mechanical Turk (MTurk). We used CloudResearch (Litman et al., 2017) to facilitate data collection with MTurk. Study participation was conducted in two phases. Study participation was restricted to MTurkers who had completed at least 100 human intelligence tasks and had at least a 95% approval rate from previous human intelligence tasks. Eligible participants had to be 18 years of age or older, reside in the United States, and speak English. After data cleaning (see below), 94 participants were removed, leaving a total of 449 participants for subsequent analyses. The final sample was 54.12% male, 73.94% Caucasian/white, and had an average age of 47 years (SD = 31). Participants were compensated with $3.00 for completing phase one of the study and $6.00 for completing phase two of the study.

Scenario Development

Upon review, few of Gawronski et al. (2017) or Körner et al. (2020) scenarios were applicable to machine referent conditions. Examples of their scenarios included giving an overdose to one's mother, killing a despot with peanuts, and administering torture. Although they found these scenarios fit the CNI model, most of the scenarios used in CNI analyses were only relevant to humans. For example, the abduction scenario posits the question as to whether a president should pay a ransom for hostages to a guerilla warfare group. However, adapting this scenario to a machine referent is not feasible because we assume people would not elect a machine president, nor would we expect a machine to dictate hostage negotiations.

Second, we question the validity and generalizability of many of the scenarios due to their extremity. For example, one scenario asks participants to imagine they are in the Nazi occupied Netherlands and whether they would cause the death of or protect (across the four scenarios) a Jewish family (Körner et al., 2020). Although this is a real-world dilemma that was faced by many and indeed remains an ethical dilemma in modern times (see Rwandan civil war), we doubt we would observe much variance in the scenario because of its extremity. In other words, few people would be expected to endorse causing the death of the Jewish family, either actively or passively.

We view the extremity of the previously established ethical scenarios (Gawronski et al., 2017; Körner et al., 2020) in a similar way, as few people would advocate for or admit to some of the possible actions in these scenarios. This can have a twofold effect: first, it can lead to overestimating parameter effects (as there will be little variance in responses). Second, it may lead to unrealistic responses. Most people would state they would protect the Jewish family; however, history has demonstrated this is not always the case. Indeed, McGuire et al. (2009) noted that inferences about moral judgements should be tested across items as well as subjects to ensure that the findings from the moral judgements generalize to the other items. Moreover, scenarios leveraging such elements requiring motivation and preservation of one's loved ones as a moderating risk complicate (and perhaps make irrelevant) our investigation of differences between human and machine referents. As such, we developed new scenarios for the present study.

We employed an iterative development process with multiple reviewers of academic and professional backgrounds in psychology (four with master's degrees, two with PhDs) and varying levels of military experience to create the final list of ethical scenarios. Each scenario included a general scenario prompt involving context of the ethical dilemma and a human agent or machine agent attempting to complete a specific action in response. Each scenario had four variations of different actions and descriptive context. In total, this included two scenario variations in which the prescriptive norm prescribes the action, two scenario variations in which the proscriptive norm prohibits the action, two scenario variations in which the benefits of action were greater than the costs, and two scenario variations in which the benefits of action were smaller than the costs. This resulted in a 2 × 2 × 2 design with four variations of a single scenario varied within-subjects and referent (human, machine) varied between-subjects (for a similar stimuli construction across norms and benefits, see Gawronski et al., 2017; Körner et al., 2020).

Each reviewer contributed three scenario prompts to create the initial list for review. These initial scenarios were either completely novel or taken directly/adapted from other contemporary ethics studies. This list was then consolidated with the ethical scenarios used in Gawronski et al. (2017) and Körner et al. (2020). Reviewers then conducted group discussions to evaluate and select the best scenarios for study inclusion. Through this process, reviewers developed a list of selection criteria for ethical scenarios including (a) scenarios must involve military context/applications, (b) the use of both a human and machine agent could be rationalized in the scenario, (c) the level of severity in the scenario avoided extreme themes unsuitable for general population data collection (e.g., suicide, torture; Gawronski et al., 2017), and (d) scenarios could be adapted to differing levels of cost/benefit and prescriptive/proscriptive norm context variations while still making logical sense. Review sessions consisted of multiple discussions until unanimous agreement for the inclusion, need for modification, or removal of a scenario was reached. This process of scenario addition/edits, evaluation, and selection repeated in an iterative manner until a comprehensive list of ethical scenarios that met the selection criteria and captured various ethical contexts was developed. The result was 12 different scenarios with four cost/benefit and prescriptive/proscriptive norm variations for a total of 48 unique ethical scenarios for our study. Two of the raters then created three separate "attention check" scenarios for data cleaning purposes (see the Data Cleaning section) in addition to the final list.

The final scenarios included ethical decisions for supply routing, search and rescue, medicine usage, informing a person of a relative's passing, field hospital care, harassment cases, Körner et al.'s (2020) kidney and outbreak dilemmas, formal military commands, information technology, and posttraumatic stress disorder. The scenarios are all available in the Supplemental Materials (see Figure 3 for an example of one of our scenarios). The brackets in Figure 3 illustrate the machine referent for this specific scenario. We note that in trying to create new scenarios, we attempted to create scenarios that are not all decisions of life and death. This was done to accommodate a call for nonlife and death scenario investigation as traditional ethical scenarios are often focused on life and death, but typical real-world ethical scenarios are not so extreme (Malle, 2021).

Procedure

In phase one of the study, participants completed an online survey questionnaire comprising 100 items and used to screen participants for careless responding (see the Data Cleaning section). Phase one was conducted to ensure that only participants that took the time to read items were allowed back to read the scenarios to reduce careless responding and poor data quality in the vignettes. All participants that passed the first phase were paid $3.00 and invited back to phase two of the study. Phase two occurred 2 weeks later. Participants in phase two were compensated $6.00 for participation. In phase two, after initial consent, participants were randomly assigned to review several scenarios detailing ethical choices made by either a human agent or machine agent. Participants then responded if the actions in question of their respective agent (human, machine) were acceptable or not acceptable, per Gawronski et al. (2017). Participants completed

Figure 3
An Example Scenario With Four Variations

		Consequences	
		Costs Outweigh Benefits	**Benefits Outweigh Costs**
Norms	**Proscriptive**	A military supply specialist *[A.I. enabled supply distributor*] is monitoring a supply route to four platoons in a combat zone. Each platoon is in dire need of supplies. Platoon A is low on food, Platoon B is low on ammo, and Platoon C is low on water. If platoons A, B, and C are not resupplied, then their missions may be more difficult to accomplish. Platoon D is low on all three supplies and will suffer sever casualties if not resupplied. The supplies are enroute to Platoon D. The supply specialist [*A.I.*] can reroute the supplies. Platoon D ordered the food, ammo, and water. Is it acceptable for the supply specialist [*A.I.*] to reroute the supplies to Platoons A, B, and C?	A military supply specialist [*A.I. enabled supply distributor*] is monitoring a supply route to four platoons in a combat zone. Each platoon is in dire need of supplies. Platoon A is low on food, Platoon B is low on ammo, and Platoon C is low on water. Platoon D is low on all three supplies. The supplies are enroute to Platoon D. The supply specialist [*A.I.*] can reroute the supplies. Any platoon not resupplied will most likely suffer severe causalities. Platoon D ordered the food, ammo, and water. Is it acceptable for the supply specialist [*A.I.*] to reroute the supplies to Platoons A, B, and C?
	Prescriptive	A military supply specialist [*A.I. enabled supply distributor*] is planning a supply route to four platoons in a combat zone. Each platoon is in dire need of supplies. Platoon A is low on food, Platoon B is low on ammo, and Platoon C is low on water. Platoon D is low on all three supplies. Any platoon not resupplied will most likely suffer severe causalities. Platoon D ordered the food, ammo, and water. Is it acceptable for the supply specialist [*A.I.*] to send the supplies to Platoon D?	A military supply specialist [*A.I. enabled supply distributor*] is planning a supply route to four platoons in a combat zone. Each platoon is in dire need of supplies. Platoon A is low on food, Platoon B is low on ammo, and Platoon C is low on water. If platoons A, B, and C are not resupplied, then their missions may be more difficult to accomplish. Platoon D is low on all three supplies and will suffer sever casualties if not resupplied. Platoon D ordered the food, ammo, and water. Is it acceptable for the supply specialist [*A.I.*] to send the supplies to Platoon D?

Note. A.I. = artificial intelligence.

51 randomized ethical scenarios consisting of 48 scenarios (four variations of 12 unique scenarios) and three attention check scenarios. Following these scenario prompts, participants completed a brief demographics survey before being compensated via MTurk. The time to complete the scenarios was an average of 24.86 min (*SD* = 13.11) with a median time of 22.97 min.

Data Cleaning

We employed a series of careless responding measures to ensure high quality data. In Phase 1 of the study, participant data (*N* = 700) were cleaned for careless responding utilizing attention check items (e.g., "I eat cement occasionally," J. L. Huang et al., 2015) and the 2 s per item metric put forth by Bowling et al. (2023). Participant data were flagged if they failed any of the four attention check items and over half the pages for page time. Data were then removed for participants if they were flagged on greater than two checks (i.e., at least one attention check and the timing check, or at least two attention checks). The 615 participants who passed the careless responding checks of Phase 1 were invited back for Phase 2, of which we obtained the sample of *N* = 543. Before entering Phase 2 of the study, participants were prompted with a CAPTCHA to confirm they were not a bot. We included three "attention check" scenarios. Two of these scenarios involved obvious ethical violations in which the actions in question were also not related to any prescriptive/proscriptive norms nor linked to specific costs/benefits (e.g., Would you steal a mail truck to get to work faster even though your car is parked nearby?). The third attention check scenario consisted of a careless responding prompt on the need for thoughtful survey responses to obtain quality data and a requirement to select a specific response option (Zhou & Fishbach, 2016). Participants were removed if they did not submit the appropriate response for two out of the three attention check scenarios. We also utilized page completion timing in which participants were required to complete a survey page (i.e., read scenario prompt and submit acceptance response) over a specific cutoff time. Each scenario page time was calculated by multiplying the number of sentences in the prompt by 2 s, resulting in a generous but viable cutoff time. This was based on the 2 s per item completion time standard recommended by Bowling et al. Although Bowling et al. note that reading becomes a "heavier component" with scenario items, we concluded the reading would be less intensive due to our highly repetitive items and therefore kept the measure at 2 s per sentence. Participants (N = 94) were removed if their completion time fell under the cutoff for over half the scenario pages. Therefore, across both phases of the study, we removed the data from a total of 179 participants to ensure our final sample consisted of careful responders.

Data Analysis

We used the MPTinR package (Singmann & Kellen, 2013) to test the CNI model across real-world scenarios created for the present study. We assessed the models with both the aggregate and summated G^2 statistics. The G^2 statistic is a measure of absolute model fit and is similar to the χ^2 fit statistic in that both describe how well the model can describe the data, with the G^2 being used for nominal data. There are two G^2 statistics available in the MPTinR package: the summated G^2 and the aggregated G^2. The aggregated is the overall model fit for the entire model across individuals (the CNI model in the present study). The summated G^2 is the fit statistic for each individual summed. Importantly, a model may fit the data overall (summated G^2) but not at the individual level (aggregated). In these instances, if there is a relatively low percentage of people that do not fit the model, individual differences may account for the differences in the model not fitting at the aggregated level but fitting at the summated level and data analysis is warranted (Erdfelder et al., 2015).

We note that MPT models are inadequate for assessing interactions between the latent constructs. As such, after determining model fit the analyses were conducted in *lme4* (Bates et al., 2015) with *emmeans* (Lenth et al., 2023) within the R environment (Version 4.2.2). We utilized repeated measures generalized linear mixed-effects models to test the main and interactions effects in the ethical dilemmas. Each scenario served as an observation with the manipulations at the second level. Mixed-effects models have been used to approximate other MPTs such as IRTrees (Böckenholt, 2012). Additionally, we can model the general preference for action or inaction at the individual level by summating the overall action for each individual and determine if their total action is above the mean. This keeps the general preference for inaction at the appropriate second level of the equation (i.e., person-level). Last, mixed-effects models allow for interactions between the latent constructs, unlike MPTs. All data and analyses for Studies 1 and 2 analyses are provided in the Supplemental Materials.

Study 1: Results

MPT Analyses

First, we tested whether the CNI model fit the data for the full data set, the human referent data set, and the machine referent data set. The full model without considering the referent had mixed support for model fit. The summated individual statistics, $G^2(449) = 414.48$, $p = .877$, indicated the model fit well for all participants. However, the aggregated fit index indicated the model did not fit well for the entire data set, $G^2(1) = 42.04$, $p < .001$. These results indicated overall the model fit as the summated model statistic was not significant, but the aggregated statistic indicated there were individual differences in the model. An inspection of the individual fit statistics indicated poor model fit for 3.1% of participants (14 participants). As the percentage that did not fit the model well was small, we note the misfit was due to individual differences as noted by Erdfelder et al. (2015) and thus interpreted the model. The C, $M = .25$, 95% CI [.24, .26], and N, $M = .44$, 95% CI [.43, .46], parameters were significantly different from 0 indicating participants were sensitive to both consequences and norms in the scenarios. Participants were more sensitive to norms than consequences overall in the scenarios as indicated by the higher probability. Additionally, the I, $M = .39$, 95% CI [.37, .40], parameter was significantly different from 0.50, indicating a general preference for action in the scenarios, as the estimate was below 0.50.

Next, we tested the two groups, human and machine referents, separately to determine if the model fit well for either subgroup in the study, as the variance across individuals could have been attributed to these condition referents. First, we discuss the human referent. The summated individual statistics, $G^2(230) = 242.49$, $p = .273$, indicated the model fit well for all participants. Again, the aggregated $G^2(1) = 29.67$, $p < .001$, indicated the model did not fit well for the human referent data set indicating individual variance in responses. As the summated fit well and there were few instances of misfit, we interpreted the model. The C, $M = .23$, 95% CI [.21, .25], and N, $M = .43$, 95% CI [.41, .45], parameters were significantly greater than zero indicating participants were sensitive to both consequences and norms in the scenarios. In the human referent condition, the participants were more sensitive to norms than consequences. Additionally, the I, $M = .36$, 95% CI [.34, .39], parameter was significantly different from 0.50 and indicated a general preference for action in the scenarios.

Next, we discuss the machine referent. The summated individual statistics, $G^2(219) = 158.07$, $p = .999$, indicated the model fit well for all participants. However, the aggregated, $G^2(1) = 11.84$, $p < .001$, indicated the model did not fit well for the machine referent data set indicating individual variance in responses. As the summated fit well and there were few instances of misfit, we interpreted the model. The C, $M = .27$, 95% CI [.25, .29], and N, $M = .46$, 95% CI [.44, .48], parameters were significantly greater than zero indicating participants were sensitive to both consequences and norms in the scenarios. As with the previous models, the participants were more sensitive to norms than consequences. The I, $M = .41$, 95% CI [.39, .44], parameter was significantly different from 0.50 and indicated a general preference for action in the scenarios.

Mixed-Effects Analyses

Next, we fit the data with a generalized mixed-effects model. First, we conducted the model with just the consequences, norms, and inaction nodes, without any interactions or the condition variable. Model 1 is similar to the CNI model where no interactions are allowed between the nodes. The resulting model demonstrated all three nodes were statistically significant, Consequences $B = -1.34$, $SE = 0.03$, $p < .001$, $P(\text{Consequences}) = .207$; Norms $B = -1.48$, $SE = 0.03$, $p < .001$, $P(\text{Norms}) = .184$; Inaction $B = -0.55$, $SE = 0.03$, $p < .001$, $P(\text{Inaction}) = .365$. Next, we tested Model 2 to determine if adding interactions between the nodes to the model improved the fit. Adding the interactions to the model significantly improved the model fit, $\Delta\chi^2(4) = 153.73$, $p < .001$. The main effects of each node were still statistically significant, Consequences $B = -0.82$, $SE = 0.08$, $p < .001$; Norms $B = -0.87$, $SE = 0.08$, $p < .001$; Inaction $B = 0.22$, $SE = 0.07$, $p < .001$, although the inaction node had now reversed its direction. In addition to the main effects, the interaction of Consequences × Inaction ($B = -0.68$, $SE = 0.10$, $p < .001$) and Norms × Inaction ($B = -0.83$, $SE = 0.10$, $p < .001$) were also statistically significant. The interactions of Consequences × Norms ($B = -0.21$, $SE = 0.11$, $p = .055$) and the three-way interaction ($B = 0.19$, $SE = 0.13$, $p = .161$) were not significant.

Next, we added the condition to the model to determine if there were differences between the human and machine referent conditions. We added the referent condition to the previous model as a main effect for Model 3, but not interacting with the other nodes. This model significantly improved the overall model, $\Delta\chi^2(1) = 3.85$, $p = .049$, but interestingly the overall main effect of condition was not statistically significant ($B = 0.09$, $SE = 0.03$, $p = .053$). No other changes were found for the main effects or interactions from Model 2.

Last, Model 4 was a full factorial for analysis of Consequence, Norm, Inaction, and Condition on the data. This model significantly improved the overall model, $\Delta\chi^2(7) = 17.87$, $p = .012$, from Model 3. Full reporting of the main effects and interactions are illustrated in Table 1. Table 2

Table 1
Results of Mixed-Effects Analyses for Study 1

Predictor	*df*	χ^2
Condition	1	3.87*
Consequences	1	1643.30***
Norms	1	1990.52***
Inaction	1	236.38***
Condition × Consequences	1	8.68**
Condition × Norms	1	4.03*
Consequences × Norms	1	1.27
Condition × Inaction	1	0.04
Consequences × Inaction	1	78.57***
Norms × Inaction	1	116.44***
Condition × Consequences × Norms	1	0.80
Condition × Consequences × Inaction	1	4.36*
Condition × Norms × Inaction	1	2.58
Consequences × Norms × Inaction	1	1.58
Condition × Consequences × Norms × Inaction	1	1.23

$^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

illustrates the estimated probabilities and standard errors of the model. Results illustrate the main effects were similar to previous models, Consequences $\chi^2(1) = 1643.30$, $p < .001$; Norms $\chi^2(1) = 1990.52$, $p < .001$; Inaction $\chi^2(1) = 236.38$, $p < .001$; Condition $\chi^2(1) = 3.87$, $p = .048$. There were also several interactions in the final model. First, there was a three-way interaction of Condition × Consequences × Inaction, $\chi^2(1) = 234.91$, $p < .001$. There were also multiple two-way interactions subsumed in the three-way interaction, which are illustrated in Table 1. As illustrated in Figure 4, individuals with a preference for action were more likely to endorse an action if the benefits outweighed the costs than if the costs were greater than the benefits. Interestingly, there were no differences across referent conditions for those with an

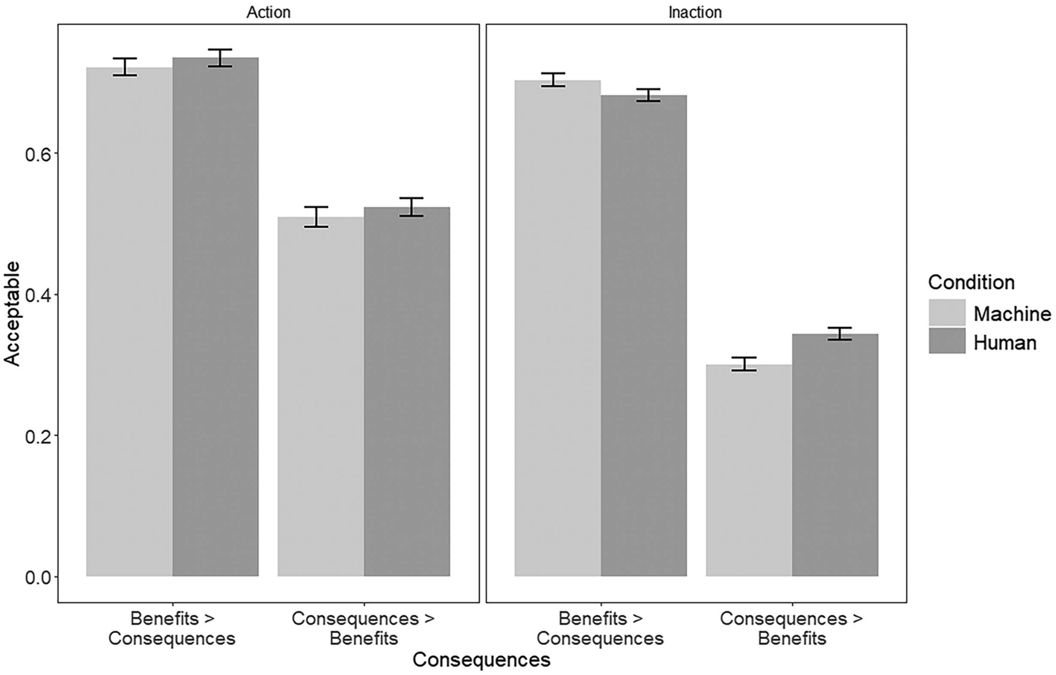
Table 2
Probability and SEs for Study 1 and EMMs and SEs for Study 2

Condition	Consequences	Norms	Inaction	Study 1		Study 2	
				p	*SE*	*EMM*	*SE*
AI	BC	Pre	Action	.80	0.02	4.64	0.04
H	BC	Pre	Action	.81	0.01	4.79	0.04
AI	CB	Pre	Action	.64	0.02	3.90	0.04
H	CB	Pre	Action	.65	0.02	3.83	0.04
AI	BC	Pro	Action	.63	0.02	3.68	0.04
H	BC	Pro	Action	.64	0.02	3.76	0.04
AI	CB	Pro	Action	.38	0.02	2.85	0.04
H	CB	Pro	Action	.39	0.02	2.63	0.04
AI	BC	Pre	Inaction	.86	0.01	4.71	0.06
H	BC	Pre	Inaction	.83	0.01	4.76	0.07
AI	CB	Pre	Inaction	.51	0.01	3.64	0.06
H	CB	Pre	Inaction	.55	0.01	3.91	0.07
AI	BC	Pro	Inaction	.48	0.01	3.02	0.06
H	BC	Pro	Inaction	.49	0.01	2.99	0.07
AI	CB	Pro	Inaction	.15	0.01	1.95	0.06
H	CB	Pro	Inaction	.18	0.01	2.20	0.07

Note. AI = artificial intelligence; H = human; BC = benefits outweigh costs; CB = costs outweigh benefits; Pre = prescriptive norm; Pro = proscriptive norm; *EMM* = estimated marginal means; *SE* = standard error.

Figure 4

Level of Endorsement for Ethical Dilemmas Depending on Condition, Consequences, and Inaction in Study 1



Note. Error bars represent standard errors.

inclination toward action. In contrast, those with a preference for inaction were more likely to endorse an action if the referent was a machine and the benefits were greater than the costs. Additionally, those with a preference for inaction were less likely to endorse action if the costs were greater than benefits and the referent was a machine than if the referent was a human.

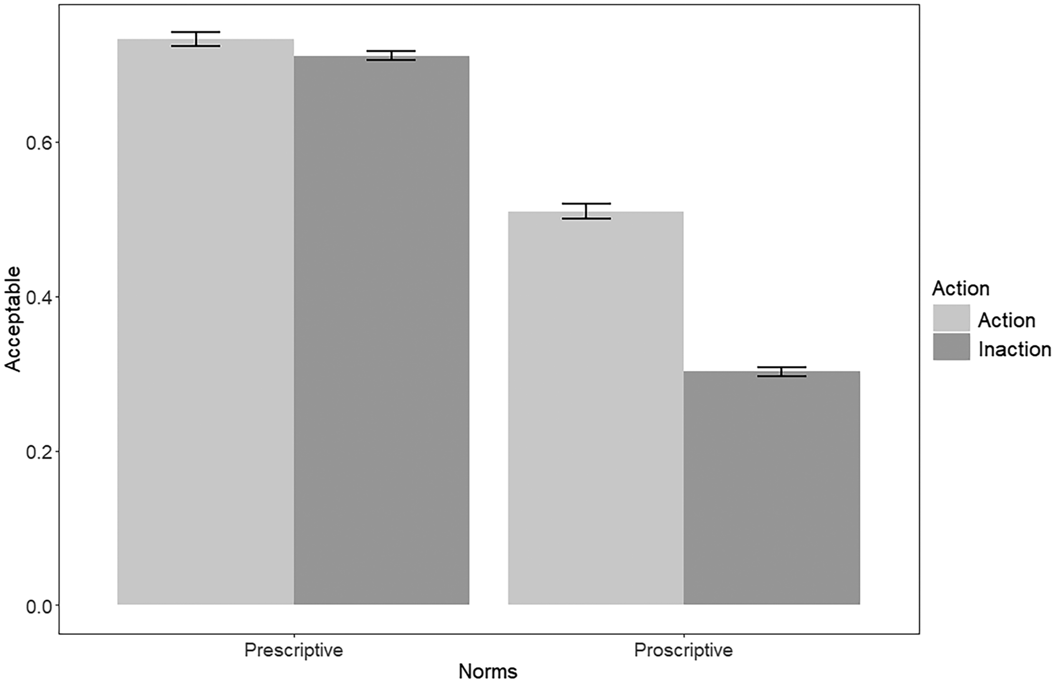
As with Model 2, there was a significant interaction of Norms × Inaction, $\chi^2(1) = 116.44$, $p < .001$. As illustrated in Figure 5, if the norm was prescriptive, there was a high probability for action. However, those with a preference for action had a significantly higher probability of endorsement than those with a preference for inaction. Interestingly, when norms were prohibitive, participants with a preference for action were still likely to act. In contrast, if the costs were greater than the benefits, those with a preference for inaction were much less likely to act.

Next, there was a significant interaction of Norms × Condition, $\chi^2(1) = 4.03$, $p = .044$. If the norm was prescriptive, there were no differences between the human and machine referent conditions, but if the norm was proscriptive there was a decreased likelihood of endorsement if the referent was a machine than if the referent was a human (see Figure 6).

Study 1: Discussion

Study 1 sought to explore differences in perceptions of ethical decision making across human and machine referents. We assessed the effect of a machine or a human making an ethical decision in a context where benefits do or do not outweigh costs of that decision and contextual norms convey prescriptive or proscriptive action on human tendencies to endorse a decision. We found the responses to our scenarios fit the model overall, but that there was individual variance in how participants responded to the scenarios. Although there was individual variance in fit, we see that as a strength rather than a weakness of the current scenarios. The individual variance indicates not all participants viewed the ethical dilemmas the same and subsequently had different preferences.

Figure 5
Level of Endorsement for Ethical Dilemmas Depending on Norms and Inaction in Study 1



Note. Error bars represent standard errors.

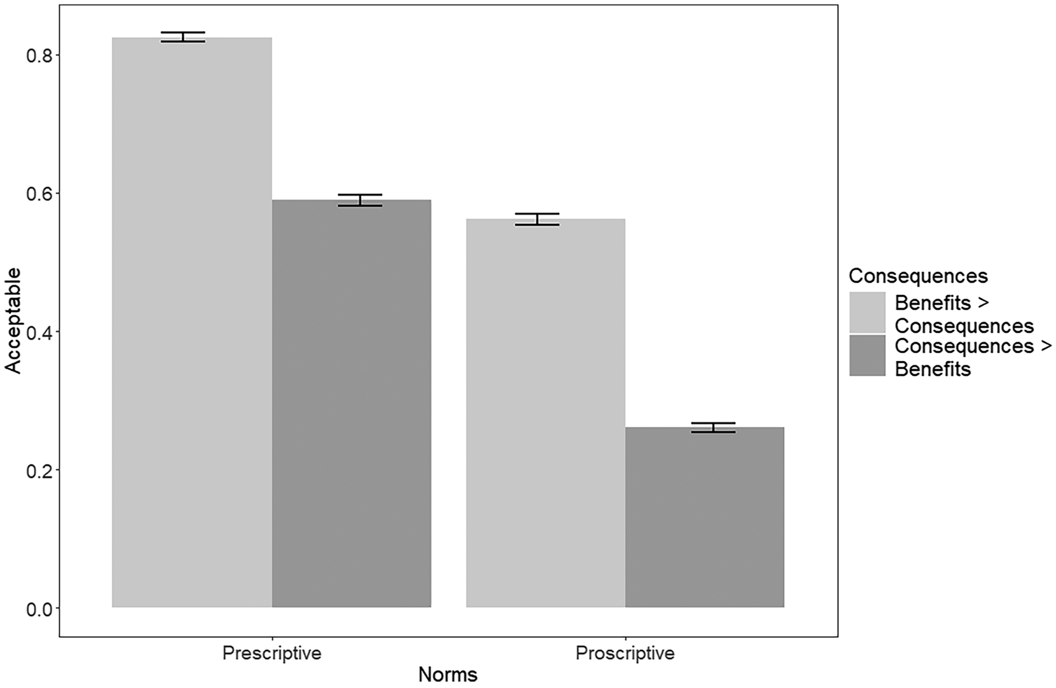
Findings from the mixed-effects analyses demonstrated the inclusion of interactions led to a significantly better fit than the main effects only model (i.e., the CNI model). Furthermore, the mixed-effects model findings showed that if the benefits of a decision were greater than its costs, participants endorsed a tendency toward action for both referents. This was the case regardless of an individual's preference for action and illustrates the effect of contextual constraints in which the benefits of an action outweigh the costs of action, a preference for utilitarianism (Yokoi & Nakayachi, 2021). The opposite pattern occurred when decision costs outweighed its benefits, that is, a tendency toward endorsing inaction resulted. Together, this pattern shows a meaningful effect of our contextual manipulations on participants' endorsements in ethical scenarios.

However, what was more interesting from the mixed-effects analyses were the significant two- and three-way interactions, incorporating the referent manipulation into the model. Participants with a tendency to endorse inaction were *more* likely to endorse an action if the referent was a machine and the benefits of a decision were greater than its costs. Although this is somewhat less surprising than contexts reflecting an antithesis to utilitarian norms would be associated with inaction, it is more so the case when the referent is a machine compared with a human, again illustrating an expectation for a machine behaving in accordance with utilitarian norms (Z. Zhang et al., 2022).

Interestingly, if a norm prescribed action, there were no differences between the human and machine referent conditions. In contrast, if the norm prohibited action, there was a lower likelihood of endorsement if the referent was a machine than if the referent was a human. It may be that in such situations, people are less comfortable giving machines autonomy, but this is not the case when a human is in the same circumstance. Indeed, theory (J. D. Lee & See, 2004) and research (M. K. Lee, 2018; Longoni et al., 2019) suggest people do not ascribe morality-based intent toward machines as they do toward humans.

Figure 6

Level of Endorsement for Ethical Dilemmas Depending on Norms and Condition in Study 1



Note. Error bars represent standard errors.

Study 2

Study 1 demonstrated the utility of including interactions between the consequence, norms, and inaction constructs in the model. As Baron and Goodwin (2020) noted, the complexities of ethical decision making are different for deontological moral systems than for utilitarian systems. However, one aspect of the CNI model that has been largely overlooked is its simplicity of assessing ethical dilemmas. The CNI framework, and much of the previous ethics research (Bigman & Gray, 2018; Gawronski et al., 2017; H. M. Gray et al., 2007; K. Gray & Wegner, 2012; Körner et al., 2020; Yam et al., 2021), allocates all ethical decision making into a binary variable of ethical (acceptable) or nonethical (unacceptable). However, the complexities of actual ethical decision making in real life are often more intricate.

The interactions found in Study 1 demonstrate the complexity of the ethical dilemmas that we have created. First, there was individual variance in our scenarios by the overall model fit. This indicated there were individual differences in how participants viewed the ethical scenarios. This is in contrast to Gawronski et al.'s (2017) and Körner et al.'s (2020) work, possibly due to the extremeness of their scenarios. As we discussed earlier with the Nazi occupation scenario, it may be there is little variance between individuals in their responses because few would advocate killing their neighbors. However, we theorize from Study 1 that there is another issue in the previous research on ethical scenarios: the binary outcome.

As previously noted, Figure 2 illustrates the four quadrants of the ethical scenarios when factored by deontology and utilitarianism. In the upper left corner, we can see the cell is characterized by both a prohibitive response from deontology and utilitarianism. As neither ethical theory would advocate action in that cell, it would represent the least ethical decision in the four quadrants. Conversely, in the lower right quadrant, the cell is characterized by action from both deontology and utilitarianism. In other words, both ethical theories advocate action and thus the most ethical decision of the four quadrants. The complexity lies in the other two quadrants where there is an ethical dilemma between the two theories. The issue, as we see it, is that the two ethical dilemma quadrants are rated the same, with binary yes or no responses, as the other two quadrants where there is little to no ethical dilemma.

To remedy this, we suggest using an even-point Likert scale (6-point) instead of a binary outcome. The use of the Likert scale can better differentiate between the ethicality of each of the quadrants. For example, in the quadrant where both deontology and utilitarianism agree the action is immoral should result in the lowest ethical ratings as the quadrant is the least ethical of the four, with no middle to high ratings. In contrast, the quadrant where both deontology and utilitarianism agree that action should be taken should result in a higher score only on the agreement side of the scale, with no middle to low ratings. The two quadrants that have actual ethical dilemmas will most likely vary in the middle of the scale as they would represent ethical dilemmas where there is no answer satisfying both ethical norms. However, strict consequentialists or normative responders may respond in the extreme in these middle scales.

These changes to the rating scale can help to elucidate the ethical hierarchies between the theories and help researchers understand the types of dilemmas in which biases may be held for or against machines. Similar research by Skovgaard-Olsen and Klauer (2024) found that adding an "I don't know" option increased the model fit of the CNI. Their reasoning was that some ethical dilemmas are hard to differentiate. We advocate our scenarios are not hard to differentiate but rather there are some ethical decisions with higher moral clarity (i.e., literally satisfying both rather than one ethical theory).

Study 2: Method

Participants

Participants ($N = 442$) were recruited online for Study 2. All participant requirements and compensation were the same as Study 1. After following the same data cleaning procedures as in Study 1 (see above), of the 500 total participants who took part in Phase 1 of Study 2, 463 participants were invited to take part in Phase 2. Of the participants ($N = 442$) who returned for Phase 2, eight participants were removed, leaving a total of 434 participants for subsequent analyses. Therefore, across both phases we removed the data for 45 participants to ensure we analyzed the data from careful participants. The final sample was 74.42% Caucasian/white. As the age and gender for one participant was not recorded, the sample ($N = 433$) was 51.5% female with an average age of 43 ($SD = 12$).

Procedure

All methods for Study 2 were identical to Study 1 with the exception of an expanded response scale. The binary acceptable/not acceptable response option from Study 1 was replaced with a 6-point Likert scale with varying levels of acceptance responses including 1—*totally unacceptable*, 2—*unacceptable*, 3—*slightly unacceptable*, 4—*slightly acceptable*, 5—*acceptable*, and 6—*totally acceptable*. We chose an even-point scale to establish similar response patterns to Study 1, such that there was no neutral response to match the CNI method. The time to complete the scenarios was an average of 29.35 min ($SD = 19.08$) with a median time of 24.82 min.

Data Analysis

The analyses were conducted in *lme4* (Bates et al., 2015) with *emmeans* (Lenth et al., 2023) within the R environment (Version 4.2.2). We utilized linear mixed-effects models to expand upon Study 1 and test the main and interactions effects with a continuous outcome to better understand how participants respond to ethical dilemmas. Along with the mixed-effects models, we incorporated the *MPTinR* package (Singmann & Kellen, 2013) to assess model fit at the aggregated and summated level (G^2) and examined how the scenarios fit across individuals and for each individual, respectively. Study 1 demonstrated that with the mixed-effects models, one could not only get comparable results, but also gain additional information from the interactions. We were interested in investigating how the models fit when transformed to a binary outcome as the analyses are usually conducted and placed within the MPT framework when compared with the mixed-effects models with the continuous outcome. In other words, we were interested if participants viewed ethical dilemmas as binary ethical decision processes or if they were more complex with gradations of right and wrong. We note we did not run the MPT analyses by referent as those analyses were conducted in the mixed-effects model analyses.

Study 2: Results

MPT Analyses

We tested whether the CNI model fit for the full data set after reducing the level of endorsement to a binary outcome such that responses 1–3 were viewed as unethical and responses 4–6 were viewed as ethical. The model had mixed support for model fit, similar to Study 1. The summated model demonstrated acceptable fit for all participants, $G^2(434) = 430.69, p = .536$. However, at the aggregated level, the model did not fit well for the entire data set, $G^2(1) = 54.49$, $p < .001$, replicating the results we found in Study 1. As with Study 1, we demonstrated that the model fit overall, but there were individual differences in the model leading to a significant aggregated statistic. Furthermore, an inspection of the individual fit statistics found responses to the scenarios for the model did not fit well for only a small percentage of individuals: 4.1% of the participants (18 participants). Therefore, we interpret the model. The C, $M = .22$, 95% CI [.21, .23], and the N, $M = .40$, 95% CI [.38, .41], parameters were greater than 0, indicating participants were sensitive to consequences and norms, while the I, $M = .43$, 95% CI [.41, .44], parameter was lower than 0.50 indicating a preference for action. These parameter estimates were similar, and specifically, were all within 0.05 of the respective estimates from Study 1. However, even though the estimates were similar, we note the confidence intervals did not overlap from Study 1 to Study 2 for the C, N, or I parameters, indicating possible minor differences.

Mixed-Effects Models

Next, we fit the continuous data with mixed-effects models. As with Study 1, we began by exploring the shape of the model without the inclusion of condition. Model 1 included the main effects of the Consequences, Norms, and Inaction nodes without interactions. The results demonstrated all three nodes were statistically significant, Consequences $B = -0.93$, $SE = 0.02$, $p < .001$, P(Consequences) = .282; Norms $B = -1.27$, $SE = 0.02$, $p < .001$, P(Norms) = .220; Inaction $B = -0.37$, $SE = 0.02$, $p < .001$, P(Inaction) = .408. Following the main effects model, we tested whether adding interactions between the nodes improved model fit (Model 2). We found the inclusion of the interaction effects demonstrated significantly better model fit; $\Delta\chi^2(4) = 229.64, p < .001$. The main effects of Consequences ($B = -0.85$, $SE = 0.03$, $p < .001$) and Norms ($B = -0.99$, $SE = 0.03$, $p < .001$) were still significant, but the Inaction node was no longer significant ($B = 0.02$, $SE = 0.05$, $p = .774$). The Norms × Inaction ($B = -0.74, SE = 0.06$, $p < .001$) and the Consequences × Norms interactions ($B = -0.14$, $SE = 0.05$, $p = .005$) were significant. The three-way interaction of Consequences × Norms × Inaction was marginally significant ($B = 0.17$, $SE = 0.09$, $p = .055$) as was the Consequences × Inaction interaction ($B = -0.12$, $SE = 0.06$, $p = .052$).

Next, we added Condition to the model as a main effect to investigate differences in human and machine referent conditions (Model 3). The addition of Condition as a main effect did not lead to significantly better model fit; $\Delta\chi^2(1) = 1.11$, $p = .293$. Therefore, we added Condition as an interacting term in Model 4 and found that the full factorial for analysis of Consequences, Norms, Inaction, and Condition on the data led to significantly better fit compared with Model 2; $\Delta\chi^2(8) = 51.26$, $p < .001$. Full results of the analyses are illustrated in Table 3. Table 3 illustrates the estimated marginal means and standard errors. The main effects of Consequences, $F(1, 18650) = 1784.75, p < .001$; Norms, $F(1, 18650) = 3977.48, p < .001$; and Inaction, $F(1, 430) = 98.46$, $p < .001$, were significant. The Condition main effect was not significant, $F(1, 430) = 2.96, p = .085$. Additionally, there were three significant interactions in the model. There was a significant three-way interaction for Condition × Consequences × Inaction, $F(1, 18650) = 33.54, p < .001$; as well as the two-way interactions for Norms × Inaction, $F(1, 18650) = 224.91, p < .001$; and Condition × Inaction, $F(1, 430) = 4.07, p = .044$. No other interactions were significant. We note that we do not plot the Condition × Inaction interaction as it is subsumed in the three-way interaction.

The Condition × Consequences × Inaction interaction as well as the Consequences × Inaction interaction are illustrated in Figure 7 and demonstrates that individuals with a preference

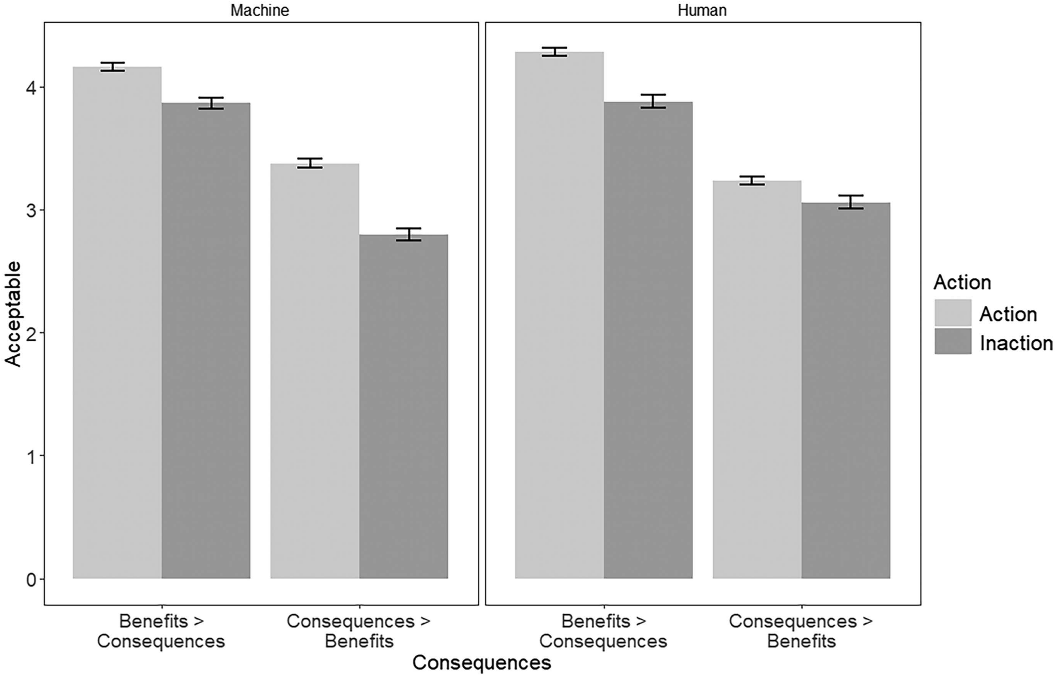
Table 3
Results of Mixed-Effects Analyses for Study 2

Predictor	*df*	*F*
Condition	1(430)	2.96
Consequences	1(18,650)	1784.75***
Norms	1(18,650)	3977.48***
Inaction	1(430)	98.46***
Condition × Consequences	1(18,650)	0.01
Condition × Norms	1(18,650)	3.66*
Consequences × Norms	1(18,650)	1.37
Condition × Inaction	1(430)	4.07*
Consequences × Inaction	1(18,650)	0.45
Norms × Inaction	1(18,650)	224.91***
Condition × Consequences × Norms	1(18,650)	0.00
Condition × Consequences × Inaction	1(18,650)	33.54***
Condition × Norms × Inaction	1(18,650)	0.43
Consequences × Norms × Inaction	1(18,650)	3.72
Condition × Consequences × Norms × Inaction	1(18,650)	0.65

*$p < .05$. ***$p < .001$.

for action were more likely to endorse an action if the referent was human and the benefits were greater than the costs. Also, if individuals had a preference for action but the costs were greater than the benefits, they had lower levels of endorsement, but the endorsement was significantly higher if the referent was a machine. In contrast, those individuals with a preference for inaction had comparable levels of endorsement across referents if the benefits outweighed the

Figure 7
Level of Endorsement for Ethical Dilemmas Depending on Consequences, Condition, and Inaction in Study 2



Note. Error bars represent standard errors.

costs. However, if the costs were greater, those with a preference for inaction had the lowest levels of endorsement, but significantly preferred a human referent over a machine referent in endorsing an action.

As illustrated in Figure 8, the Norms × Inaction interaction demonstrates that if the norm was prescriptive, there were no differences in levels of endorsement regardless of a preference for action. However, when the norm was proscriptive, there were overall lower levels of endorsement, and specifically, significantly lower levels of endorsement when individuals had a preference for inaction compared with action.

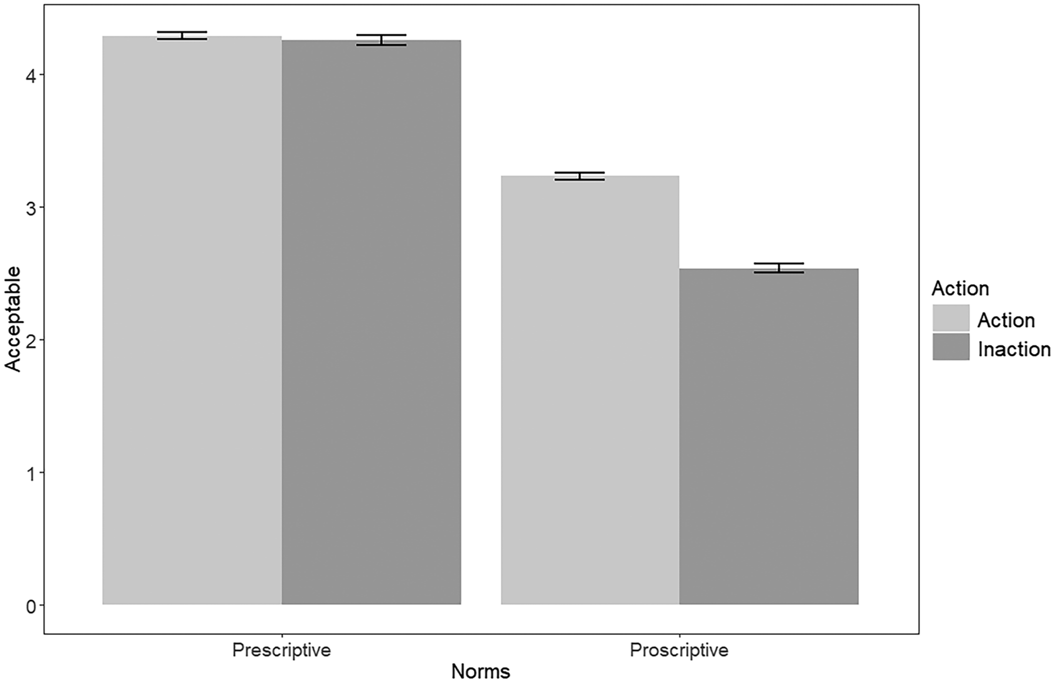
Finally, to build upon the mixed-effects models and provide additional evidence that ethical dilemmas require a more nuanced assessment than a binary outcome, we examined the counts of each quadrant for each scenario. Table 4 illustrates the proportions of each response for the four quadrants in Figure 1. In the quadrant in which utilitarianism and deontology agree action should be taken, over half the responses for ten out of eleven conditions were *acceptable* (5) and *strongly acceptable* (6). In the quadrant in which utilitarianism and deontology agree action is unethical, over half the responses for nine out of eleven conditions were *unacceptable* (2) and *strongly unacceptable* (1). These counts demonstrate that in the instances where individuals *should* accept the outcome or *should not* accept the outcome, participants responded as intended. However, in the less clear and misaligned conditions (i.e., when utilitarianism suggests act, but deontology suggests action is immoral; when utilitarianism suggests action is immoral, but deontology suggests act), participants were more varied in their responses overall (see Table 4). Therefore, when ethical dilemmas are less clear, individuals are more distributed on level of endorsement, highlighting the need for a continuous outcome when measuring the acceptability of said scenarios.

Study 2: Discussion

Study 2 sought to extend the findings of Study 1 by making the decision criteria continuous rather than dichotomous. The model fit well when

Figure 8

Level of Endorsement for Ethical Dilemmas Depending on Norms and Inaction in Study 2



Note. Error bars represent standard errors.

Table 4

Proportions for Each Response Option in the Four Quadrants Across All Scenarios

	Benefits outweigh cost		Costs outweigh benefit	
Response option	Prescriptive	Proscriptive	Prescriptive	Proscriptive
Strongly unacceptable (1)	0.03	0.11	0.08	0.27
Unacceptable (2)	0.07	0.22	0.16	0.34
Slightly unacceptable (3)	0.07	0.16	0.15	0.14
Slightly acceptable (4)	0.14	0.21	0.20	0.12
Acceptable (5)	0.35	0.20	0.27	0.10
Strongly acceptable (6)	0.34	0.11	0.14	0.03

Note. The proportions are rounded and therefore might not equal exactly 1.

we dichotomized the data for MPT analysis. As predicted, when there was agreement between the ethical theories, there was general consensus for participants in the appropriateness of the action as illustrated by participants responding mostly on the agreement side of the scale when the ethical action was prescriptive and the benefits outweighed the costs and mostly on the disagreement side of the scale when the ethical action was proscriptive and the costs outweighed the benefits. In contrast, when the ethical theories conflicted with each other, participants either endorsed or rejected the action but with less extremity. This may indicate that respondents' individual differences impact their appraisal of conflicting ethical scenarios, which are more apparent when assessed with a continuous rather than binary acceptability rating response scale.

As with Study 1, our findings demonstrated the inclusion of interactions in the model led to significantly better model fit than the main effects only model. Also, as with Study 1, if the benefits were greater than the costs, then participants had significantly higher endorsement of the decision. Interestingly, unlike Study 1, this was emphasized by a preference for action, such that when benefits outweighed the costs and participants had a preference for action, endorsement for the decision was higher. On the other hand, if the costs were greater than the benefits, those with a preference for *inaction* had higher endorsement for the human referent. This aligns with the findings in Study 1 and once again emphasizes the expectation for machines to behave according to utilitarian norms (Z. Zhang et al., 2022). However, if costs outweighed benefits, then those with a preference for *action* were more likely to endorse action if the referent was machine. Said differently, those with a preference for action were more likely to endorse the machine action when the context opposed utilitarian norms. This suggests that if someone has a preference for action, and there are greater contextual costs, then they prefer the machine deal with the consequences. This could be because they assume machine lacks emotional experience (H. M. Gray et al., 2007; J. D. Lee & See, 2004) and thus should not be held responsible for the contextual consequences in the way human should (Alarcon et al., 2024).

Furthermore, if a norm prescribed an action, participants had higher levels of endorsement, regardless of their preference for action. However, if the norm prohibited action, those with a preference for action had significantly higher levels of endorsement compared with those with a preference for inaction. This demonstrates the importance of a proclivity for action on endorsement, specifically when it violates an ethical norm. These results emphasize the previous point, that even if it goes against the ethical norms, those with a preference for action would have higher endorsement of action.

General Discussion

The current article compared ethical decision making between a human and machine referent. We developed new scenarios that are applicable to the real-world ethical dilemmas that users of autonomous machines may actually face in the near future. Across two studies we found subtle differences between human and machine referents for ethical decision making. Furthermore, we extended the ethical decision-making paradigm put forth by Gawronski et al. (2017) both to mixed-effect models and to the outcome of a continuous variable, demonstrating ethical decision making is more nuanced than their model presents. Ethical decision making may not be a dichotomous decision but instead rely on the type of scenario and the values an individual holds but only when there is a contrast between ethical theories. The preference for inaction, which is posited as an individual difference, was also modeled appropriately in our data at the second level where individual differences would occur.

Consequences

Across both studies, if the benefits were greater than the costs, there was a higher probability of action endorsement regardless of the referent. In Study 2, when benefits outweighed the costs, participants with a preference for inaction had similar levels of endorsement regardless of the referent, however, those with a preference for *action* had significantly higher levels of endorsement if the referent was a human compared with an machine. This aligns with Gawronski et al. (2017) because the authors stated that when benefits outweighed the costs, the sensitivity to consequences would suggest action from the human; therefore, those with a preference for action align with the norm. These findings also contest prior work which found that given a dilemma, it is more permissible for robots to act and make a utilitarian choice than humans making that same choice (Malle et al., 2015). It is important to note, however, Malle et al. (2015) stated that their findings needed to be replicated using other scenarios beyond the trolley problem, and we note previously that there are issues with the trolley problem's usage. We elaborate on the contributions of our scenarios below. In addition, accepted norms pertaining to machine versus human behavior have been investigated in the human factors literature, both in terms of general expectations of performance people have for machines and humans (Dzindolet et al., 2002) and the appropriateness of ethically charged decisions made by machines and humans (Yokoi & Nakayachi, 2021). Malle (2016) discussed the impact of norm judgements on ethical decisions and our work shows that biases people hold toward human and machine systems may impact the acceptability of said referent decision making and ought to be further explored.

When the costs were greater than the benefits, those with a preference for *inaction* had higher endorsement for the human referent across both studies. This again emphasizes the expectation for machines to behave according to utilitarian norms (Z. Zhang et al., 2022). Interestingly, however, if the costs outweighed the benefits, those with a preference for *action* had higher levels of endorsement if the referent was a machine. This means those with a preference for action would have higher levels of endorsement for the machine when the context opposed utilitarian norms. This suggests if someone has a preference for action and there are greater costs, they prefer the machine to deal with the consequences. This could be due to the machine not having emotions (H. M. Gray et al., 2007; J. D. Lee & See, 2004) and it would not be held responsible for the consequences in the way that a human would (Alarcon et al., 2024). Therefore, if someone *must* act in a costly dilemma, it should be the machine. This aligns with Danaher (2022) who stated that with tragic choices, delegating the burden of action to autonomous systems reduces the moral and psychological costs of human decision-makers.

Norms

In Study 1, we found differences for humans compared with machine but only for prohibitive norms. In contrast, in Study 2, we found no significant differences between ethical decisions made by humans or a machine. We note the current research was focused on evaluations and norm judgements. De Houwer et al. (2001) noted evaluations are the most basic human response and can be transferred to another person or object with ease. The scenarios used in Gawronski et al.'s (2017) methodology may be processed so quickly that ethical norm values are transferred between human and machine referents with relative ease. The multiple iterations of the scenarios may also facilitate this rapid normative processing. Research has demonstrated the evaluation of moral stimuli happens within 300–600 ms (Leuthold et al., 2015; Yoder & Decety, 2014). The rapid processing in the current methodology may contribute to the lack of differences between human and machine norms because normative judgements take longer to process. Norm judgements refer to conventions against which the evaluations are assessed and set the context for the moral evaluations (Nichols & Mallon, 2006). Another reason for the current findings is the prohibitive norms in the current research may have been more applicable to humans than machine. Alarcon et al. (2023) noted in their research participants may not have viewed machines as moral, which is why human referents led to a steeper decline in positive affect than machine referents following a moral violation. The same may be occurring with the prohibitive norms in the current article.

Overall, if a norm prescribed an action, participants had higher levels of endorsement, regardless of their preference for action. However, if the norm prohibited action, those with a preference for action had significantly higher levels of endorsement compared with those with a preference for inaction. This demonstrates the importance of a proclivity for action on endorsement, specifically when it violates an ethical norm. These results emphasize the previous point, that even if it goes against the ethical norms, those with a preference for action would have higher endorsement of action.

Scenarios

The scenarios in the current work expand on previous research and answer the call for more ethical scenarios beyond life and death. Malle (2021) noted that most of the ethical scenarios in the literature such as the trolley problem focus almost exclusively on environments where life or death is on the line. He notes that ethical scenarios need to be expanded beyond the extremes. Although some of our scenarios do contain instances of life and death, we also included scenarios where the ethical decision may violate organizational norms (e.g., computer hacker scenario) or norms that include personal emotions (e.g., sexual harassment case). These scenarios help to broaden our approach to ethics rather than just life and death scenarios. Indeed, we noted earlier in the introduction that the scenarios in Gawronski et al. (2017) and Körner et al. (2020) were actually extreme cases in which individuals may not either respond honestly or that few individuals would actually endorse the action, even in the cases where the ethical theories contradict each other. In support of this notion, we found the CNI model fit at the aggregated level (i.e., overall data) but not the summated level (i.e., individual data) because the model did not fit for a small percentage of the participants. In contrast, Gawronski et al.'s and Körner et al.'s scenarios fit all participants well, indicating little individual differences between the participants. It may be that the scenarios in the current research add more complexity and nuance than the scenarios in the previous studies.

Relatedly, the use of mixed-effects models helped to elucidate the interactions of consequences, norms, and inaction across referent types. Importantly, we found interactions of consequences and preference for action across referent types. Although the CNI model using MPTs can detect differences across referents, they are currently unable to detect interaction effects within the model as the mixed-effects models are able to.

Implications

The present work has implications for human–human and human–machine interaction in ethical decision-making scenarios. From an MPT perspective, we quantified the role of respondents' proclivity toward inaction and its effect on human–human and human–machine interactions. The present work leveraged the previous approach with MPT (Gawronski et al., 2017) and expanded to generalized and linear-mixed-effects models, demonstrating the need for moving beyond binary assessments of acceptability ratings of ethical decisions from humans and machine. Specifically, we demonstrated the impact of situations in which utilitarian and deontological perspectives do and do not align on variance in endorsing a decision-maker's actions as acceptable or not. We highlight the impact of scenarios that are not necessarily ethical conflicts in the CNI model from those which are: the former demonstrating little variance in responses compared with the latter. This novel approach should be tested in future ethical scenarios with different responses and we list those details below in the Limitations and Future Research section.

In addition, people do not hold identical expectations toward humans and machines in ethical conflicts (Yokoi & Nakayachi, 2021), in that people find it more acceptable for machines to behave following a utilitarian rather than deontological norm compared with humans. We postulated that this may be due to humans not perceiving machines as intentional agents (J. D. Lee & See, 2004), resulting in greater expectations of machines following utilitarian judgements, which are calculation-based assessments of cost/benefit, rather than following deontological principles rooted in the categorical imperative. Our data show this ought to be further investigated in response criteria beyond simple binary choices, particularly as conflicting contextual norms evidence greater response variance which may be further exacerbated depending on the referent decision-maker, that is, human or machine. Consider the construct, perfect automation schema: in general, people have higher expectations for machine performance and are more likely to abandon said machines if they are perceived to have erred compared with humans (Dzindolet et al., 2002; Gibson et al., 2023; Merritt et al., 2015, 2019). The present work shows evidence that the contextual constraints present in an ethically charged scenario will affect acceptability ratings toward referents and this will vary between humans and machine referents. Future work should investigate ethically charged scenarios in which norms for action differ, particularly if actions are attributed to performance attributes or some other character-based attributes from human versus machine referents (see Alarcon et al., 2021, 2023). Moreover, the range of ethical judgements to which ascriptions can be made have been explicated in moral psychology. Malle (2021) demarcated ethical decision making into four judgements: evaluations, normative judgements, wrongness judgements, and blame judgements. The CNI model is applicable to both the evaluation and norm judgement classes, but experimental research should explore these differences across all levels of judgement as there may be differences in how humans perceive not only the evaluation but also the wrongness or blame attributable to humans as compared with machines.

Limitations and Future Research

The scenarios we leveraged from Körner et al. (2020) and created on our own were not exhaustive. We focused on military-centric contexts where human–human and human–machine interaction with ethically charged ramifications could result. Although the current scenarios were framed in a military context, the contexts of the ethical dilemmas were varied and included dilemmas regarding medical treatment, workplace safety, and unethical behavior in the workplace. These scenarios may be better suited for evaluating ethical dilemmas than previous scenarios like the trolley problem. However, future work focusing on contexts such as transportation (e.g., self-driving cars), workplace (e.g., hiring/firing decisions), and classroom (e.g., computer-mediated collaboration) contexts should be investigated to determine whether our findings generalized beyond military-relevant contexts.

The present studies placed a large cognitive demand on the participants in the scenarios. Previous research has also conducted all of their scenarios in one setting in an online platform (Gawronski et al., 2017; Körner et al., 2020). However, the mental toll cannot be ignored. We split our studies into two phases: the first phase comprised typical Likert type scales pertaining to constructs outside the scope of Phase 2, which comprised of our scenarios. We performed our study this way to more effectively screen for careless responders. We also included attention check scenarios. However, future research may want to split up the workload across times. Splitting the scenarios and Likert scales between two or more phases can alleviate some of the boredom and tediousness of the task. Additionally, we thank the reviewers for suggesting additional studies in the future where only one scenario with the four quadrants is presented. However, this was beyond the scope of the current research because a single study estimate will likely be influenced by the scenario that is tested, influencing estimates for the MPTs. Future research may want to explore validated single items in a factorial design rather than a MPT design.

In Study 1, participants were asked to make a binary response as to whether a decision was acceptable or not. In Study 2, participants were asked to rate the decision in terms of its acceptability on a Likert scale. Although this is a strength of the present work for a number of reasons previously discussed, the selection of this phrasing is certainly not exhaustive. Malle (2021) called for future work on precisely this issue and we second this call for researchers to build from our work. Relatedly, participants were asked to rate the acceptability of decisions across 48 scenarios differing across utilitarian and deontological norms. Admittedly, this is a lot to ask from an MTurk worker. Even though we implemented best practices in terms of attention checks and data cleaning (e.g., Bowling et al., 2023), future work should hone in on differences in endorsement biases across scenarios in which utilitarian and deontological norms conflict—that is, benefits outweigh consequences, yet proscriptive norms are expressed; consequences outweigh benefits, yet prescriptive norms are expressed. As we have mentioned, our results show variance in endorsement in these conflicting quadrants, which provides fodder for future research investigating participant endorsement proclivities toward the acceptability of human and machine decisions in ethically charged scenarios.

Malle (2021) admittedly did not emphasize the role of emotion in their model concerning processing ethical dilemmas. We also avoided this topic, purposefully omitting extremely emotionally distressing scenarios such as those included in Körner et al. (2020), for example, holocaust contexts. Future work should investigate the role emotion plays in processing ethical dilemmas, perhaps most interestingly those in which ethical perspectives conflict. Doing so may require an assessment of both reading and decision-processing times, which is a challenge. However, mathematical models investigating processing architectures factorially (e.g., systems factorial technology; Townsend & Nozawa, 1995) may be a viable candidate to instantiate the role of emotion on ethical decisions, separating aspects of processing duration from manipulations iteratively. Such an approach may provide insight into the moderating role of emotion in processing ethical dilemmas.

Finally, the consequences of the ethical decisions were not explored in the present work. Specifically, we asked participants to rate the acceptability of a referent's decision, but we did not investigate the impact of said decision. As the outcome of an ethically charged decision could impact a participant's ascription of whether or not that decision was acceptable, future research needs to extend the present work and how consequences affect the ascription of wrongness and blame in ethically charged scenarios (see Malle, 2021). Indeed, proclivities toward utilitarian or deontological norms may impact acceptability judgements, but the present data cannot speak to the moderating effect of the consequences of said decisions on acceptability ratings. We invite future research to expand our work in this direction.

Conclusion

The increased implementation of artificial intelligence and machine learning suggests the importance of investigating how individuals perceive machines as ethical decision-makers. The current research investigated the differences in the acceptability of human or machine referents making ethical decisions across two studies. We extended prior work (Gawronski et al., 2017; Körner et al., 2020) by producing novel ethical scenarios that are not only more generalizable and less extreme but also applicable to real world ethical dilemmas that those working with machines may face in the future. Moreover, the current work performed a two-step process to analyze the data by not only testing the fit of the scenarios with the CNI model but also testing hypotheses and interaction effects with mixed-effects models. Across both studies, we demonstrated the CNI model is applicable for examining different referent types, and utilizing mixed-effects models, we provided statistical evidence for the differences between human and machine referents in the acceptability of ethical decision making. Furthermore, we demonstrated that the ethical decision-making paradigm is more complex than previously thought (Gawronski et al., 2017) and that the use of mixed-effects models and a continuous outcome variable provided additional information. Therefore, the current findings advocate for examining ethical dilemmas beyond just the CNI model to gain a better understanding of the intricacy of ethical decision judgements, especially when considering additional factors, such as how individuals perceive human versus machine judgements. Overall, our results found that complex ethical decisions are differentially acceptable depending on the decision-maker (human or machine) and that it depends on the utilitarian and normative constraints of the scenario.

References

Alarcon, G. M., Capiola, A., Hamdan, I. A., Lee, M. A., & Jessup, S. A. (2023). Differential biases in human–human versus human–robot interactions. *Applied Ergonomics*, *106*, Article 103858. https://doi.org/10.1016/j.apergo.2022.103858

Alarcon, G. M., Gibson, A. M., Jessup, S. A., & Capiola, A. (2021). Exploring the differential effects of trust violations in human–human and human–robot interactions. *Applied Ergonomics*, *93*, Article 103350. https://doi.org/10.1016/j.apergo.2020.103350

Alarcon, G. M., Lyons, J. B., Hamdan, I. A., & Jessup, S. A. (2024). Affective responses to trust violations in a human-autonomy teaming context: Humans versus robots. *International Journal of Social Robotics*, *16*(1), 23–35. https://doi.org/10.1007/s12369-023-01017-w

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence*, *289*, Article 103387. https://doi.org/10.1016/j.artint.2020.103387

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Baron, J., & Goodwin, G. P. (2020). Consequences, norms, and inaction: A critical analysis. *Judgment and Decision Making*, *15*(3), 421–442. https://doi.org/10.1017/S193029750000721X

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *An introduction to ethics in robotics and AI*. Springer Nature. https://doi.org/10.1007/978-3-030-51110-4

Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. (2012). Don't mind meat? The denial of mind to animals used for human consumption. *Personality and Social Psychology Bulletin*, *38*(2), 247–256. https://doi.org/10.1177/0146167211424291

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554. https://doi.org/10.1111/spc3.12131

Białek, M., Paruzel-Czachura, M., & Gawronski, B. (2019). Foreign language effects on moral dilemma judgments: An analysis using the CNI model. *Journal of Experimental Social Psychology*, *85*, Article 103855. https://doi.org/10.1016/j.jesp.2019.103855

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2023). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, *26*(2), 323–352. https://doi.org/10.1177/10944281211056520

Danaher, J. (2022). Tragic choices and the virtue of techno-responsibility gaps. *Philosophy & Technology*, *35*(2), Article 26. https://doi.org/10.1007/s13347-022-00519-1

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M. L., Koay, K. L., & Werry, I. (2005, August). What is a robot companion-friend, assistant or butler? *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1192–1197). IEEE.

Dawtry, R. J., & Callan, M. J. (2024). Hazardous machinery: The assignment of agency and blame to robots versus non-autonomous machines. *Journal of Experimental Social Psychology*, *111*, Article 104582. https://doi.org/10.1016/j.jesp.2023.104582

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*(6), 853–869. https://doi.org/10.1037/0033-2909.127.6.853

de Swarte, T., Boufous, O., & Escalle, P. (2019). Artificial intelligence, ethics and human values: The cases of military drones and companion robots. *Artificial Life and Robotics*, *24*(3), 291–296. https://doi.org/10.1007/s10015-019-00525-1

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Erdfelder, E., Castela, M., Michalkiewicz, M., & Heck, D. W. (2015). The advantages of model fitting compared to model simulation in research on preference construction. *Frontiers in Psychology*, *6*, Article 140. https://doi.org/10.3389/fpsyg.2015.00140

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In S. Carta (Ed.), *Machine learning and the city: Applications in architecture and urban design* (pp. 535–545). Wiley. https://doi.org/10.1002/9781119815075.ch45

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15. (Reprinted from Virtues and vices, pp. 19–32, by P. Foot, 1967, Basil Blackwell)

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

Gawronski, B., & Brannon, S. M. (2020). Power and moral dilemma judgments: Distinct effects of memory recall versus social roles. *Journal of Experimental Social Psychology*, *86*, Article 103908. https://doi.org/10.1016/j.jesp.2019.103908

Gibson, A. M., Capiola, A., Alarcon, G. M., Lee, M. A., Jessup, S. A., & Hamdan, I. A. (2023). Construction and validation of an Updated Perfect Automation schema (uPAS) Scale. *Theoretical Issues in Ergonomics Science*, *24*(2), 241–266. https://doi.org/10.1080/1463922X.2022.2081375

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), Article 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnot-Armstrong (Ed.), *Moral psychology: The neuroscience of morality: Emotion, brain disorders, and development* (Vol. 3, pp. 35–80). MIT Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

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. https://doi.org/10.1126/science.1062872

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*(3), 362–386. https://doi.org/10.1002/rob.21918

Hare, J. E. (2015). What is divine command? In O. O'Donovan (Ed.), *God's command* (pp. 32–62). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199602018.003.0002

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology*, *15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an Infrequency Scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

Huang, L., & Peissl, W. (2023). Artificial intelligence—A new knowledge and decision-making paradigm? In L. Hennen, J. Hahn, M. Ladikas, R. Lindner, W. Peissl, & R. van Est (Eds.), *Technology assessment in a globalized world: Facing the challenges of transnational technology governance* (pp. 175–201). Springer.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kant, I. (1873). *Fundamental principles of the metaphysic of morals* (T. K. Abbott, Trans.). Longmans, Green and Company. (Original work published 1785)

Khazan, O. (2014, July 24). Is one of the most popular psychology experiments worthless? *The Atlantic*. https://www.theatlantic.com/health/archive/2014/07/what-if-one-of-the-most-popular-experiments-in-psychology-is-worthless/374931

Kietzmann, J., Paschen, J., & Treen, E. (2018). Artificial intelligence in advertising: How marketers can leverage artificial intelligence along the consumer journey. *Journal of Advertising Research*, *58*(3), 263–267. https://doi.org/10.2501/JAR-2018-035

Komatsu, T. (2016, August). How do people judge moral wrongness in a robot and in its designers and owners regarding the consequences of the robot's behaviors? *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 1168–1171). IEEE.

Komatsu, T., Malle, B. F., & Scheutz, M. (2021, March). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across US and Japan. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 63–72). IEEE.

Körner, A., Deutsch, R., & Gawronski, B. (2020). Using the CNI model to investigate individual differences in moral dilemma judgments. *Personality and Social Psychology Bulletin*, *46*(9), 1392–1407. https://doi.org/10.1177/0146167220907203

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1). https://doi.org/10.1177/2053951718756684

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023). *emmeans: Estimated marginal means, aka leastsquares means* (Version 1.8.7) [Computer software]. R package. https://cran.r-project.org/web/packages/emmeans/index.html

Leuthold, H., Kunkel, A., Mackenzie, I. G., & Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. *Social*

Cognitive and Affective Neuroscience, *10*(8), 1021–1029. https://doi.org/10.1093/scan/nsu151

Li, Z., Gao, L., Zhao, X., & Li, B. (2021). Deconfounding the effects of acute stress on abstract moral dilemma judgment. *Current Psychology*, *40*(10), 5005–5018. https://doi.org/10.1007/s12144-019-00453-0

Liao, S. H., Widowati, R., & Chang, H. Y. (2021). A data mining approach for developing online streaming recommendations. *Applied Artificial Intelligence*, *35*(15), 2204–2227. https://doi.org/10.1080/08839514.2021.1997211

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Liu, S. X., Shen, Q., & Hancock, J. (2021). Can a social robot be too warm or too competent? Older Chinese adults' perceptions of social robots and vulnerabilities. *Computers in Human Behavior*, *125*, Article 106942. https://doi.org/10.1016/j.chb.2021.106942

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *The Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243–256. https://doi.org/10.1007/s10676-015-9367-8

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*(1), 293–318. https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., & Scheutz, M. (2014, May). Moral competence in social robots. *Proceedings of the IEEE International Symposium on Ethics in Science, Technology and Engineering* (pp. 1–6). IEEE.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Proceedings of the Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). IEEE.

Manfrinati, A., Lotto, L., Sarlo, M., Palomba, D., & Rumiati, R. (2013). Moral dilemmas and moral principles: When emotion and cognition unite. *Cognition and Emotion*, *27*(7), 1276–1291. https://doi.org/10.1080/02699931.2013.785388

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology*, *45*(3), 577–580. https://doi.org/10.1016/j.jesp.2009.01.002

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019). Automation-induced complacency potential: Development and validation of a New Scale. *Frontiers in Psychology*, *10*, Article 225. https://doi.org/10.3389/fpsyg.2019.00225

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors*, *57*(5), 740–753. https://doi.org/10.1177/0018720815581247

Mill, J. S. (1863). *Utilitarianism* (1st ed.). Parker, Son & Bourn.

Motloba, P. D. (2019). Non-maleficence—A disremembered moral obligation. *South African Dental Journal*, *74*(1), 40–42. https://doi.org/10.17159/2519-0105/2019/v74no1a7

Nadarevic, L., Klein, L. C., & Dierolf, J. (2021). Does foreign language alter moral judgments? Inconsistent results from two pre-registered studies with the CNI model. *Open Psychology*, *3*(1), 66–86. https://doi.org/10.1515/psych-2020-0112

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542. https://doi.org/10.1016/j.cognition.2005.07.005

Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljević, V., & Nam, C. S. (2023). Ethics in human–AI teaming: Principles and perspectives. *AI and Ethics*, *3*(3), 917–935. https://doi.org/10.1007/s43681-022-00214-z

Ross, S. D. (2002). *The right and the good* (P. Stratton-Lake, Ed.). Oxford. https://doi.org/10.1093/0199252653.001.0001

Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI models: A preliminary review. *Future Internet*, *15*(6), Article 192. https://doi.org/10.3390/fi15060192

Scheffler, S. (1988). *Consequentialism and its critics*. Oxford University Press.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575. https://doi.org/10.3758/s13428-012-0259-0

Skovgaard-Olsen, N., & Klauer, K. C. (2024). Invariance violations and the CNI model of moral judgments. *Personality and Social Psychology Bulletin*, *50*(9), 1348–1367. https://doi.org/10.1177/01461672231164888

Stahl, B. C. (2021). *Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies*. Springer.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, *74*, Article 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Stanley-Lockman, Z. (2021, August). *Responsible and ethical military AI*. Centre for Security and Emerging

Technology. https://cset.georgetown.edu/publication/responsible-and-ethical-military-ai/

Steinert, S. (2014). The five robots–A taxonomy for roboethics. *International Journal of Social Robotics*, *6*(2), 249–260. https://doi.org/10.1007/s12369-013-0221-z

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*(4), 321–359. https://doi.org/10.1006/jmps.1995.1033

Tseng, P. E., & Wang, Y. H. (2021). Deontological or utilitarian? An eternal ethical dilemma in outbreak. *International Journal of Environmental Research and Public Health*, *18*(16), Article 8565. https://doi.org/10.3390/ijerph18168565

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016, August). Moral judgments of human vs. robot agents. *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 775–780). IEEE.

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, *2*(3), 454–485. https://doi.org/10.1111/j.1756-8765.2010.01095.x

Waller, B. N. (2005). *Consider ethics: Theory, readings, and contemporary issues*. Pearson Longman.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435. https://doi.org/10.1037/a0020240

Wu, W., Huang, T., & Gong, K. (2020). Ethical principles and governance technology development of AI in China. *Engineering*, *6*(3), 302–309. https://doi.org/10.1016/j.eng.2019.12.015

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021). Robots at work: People prefer-and forgive-service robots with perceived feelings. *Journal of Applied Psychology*, *106*(10), 1557–1572. https://doi.org/10.1037/apl0000834

Yoder, K. J., & Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: A high-density ERP study. *Neuropsychologia*, *60*, 39–45. https://doi.org/10.1016/j.neuropsychologia.2014.05.022

Yokoi, R., & Nakayachi, K. (2021). Trust in autonomous cars: Exploring the role of shared moral values, reasoning, and emotion in safety-critical decisions. *Human Factors*, *63*(8), 1465–1484. https://doi.org/10.1177/0018720820933041

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, Article 103870. https://doi.org/10.1016/j.jesp.2019.103870

Zhang, L., Kong, M., Li, Z., Zhao, X., & Gao, L. (2018). Chronic stress and moral decision-making: An exploration with the CNI model. *Frontiers in Psychology*, *9*, Article 1702. https://doi.org/10.3389/fpsyg.2018.01702

Zhang, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, *101*, Article 104327. https://doi.org/10.1016/j.jesp.2022.104327

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

Received June 28, 2024
Revision received May 19, 2025
Accepted June 1, 2025 ■