01 Mar 2022

Perceptions of Violations by Artificial and Human Actors across Moral Foundations

Timothy Maninger

Daniel Burton Shank
Missouri University of Science and Technology, shankd@mst.edu

Recommended Citation

Maninger, T., & Shank, D. B. (2022). Perceptions of Violations by Artificial and Human Actors across Moral Foundations. *Computers in Human Behavior Reports, 5* Elsevier.
The definitive version is available at https://doi.org/10.1016/j.chbr.2021.100154

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Perceptions of violations by artificial and human actors across moral foundations

Timothy Maninger, Daniel B. Shank *

Missouri University of Science and Technology, USA

ARTICLE INFO

Keywords:
Moral foundations
Theory of blame
Artificial intelligence
Morality
Moral wrongs
Artificial agents

ABSTRACT

Artificial agents such as robots, chatbots, and artificial intelligence systems can be the perpetrators of a range of moral violations traditionally limited to human actors. This paper explores how people perceive the same moral violations differently for artificial agent and human perpetrators by addressing three research questions: How wrong are moral foundation violations by artificial agents compared to human perpetrators? Which moral foundations do artificial agents violate compared to human perpetrators? What leads to increased blame for moral foundation violations by artificial agents compared to human perpetrators? We adapt 18 human-perpetrated moral violation scenarios that differ by the moral foundation violated (harm, unfairness, betrayal, subversion, degradation, and oppression) to create 18 agent-perpetrated moral violation scenarios. Two studies compare human-perpetrated to agent-perpetrated scenarios. They reveal that agent-perpetrated violations are more often perceived as not wrong or violating a different foundation than their human counterparts. People are less likely to classify violations by artificial agents as oppression and subversion, the foundations that deal the most with group hierarchy. Finally, artificial agents are blamed less than humans across moral foundations, and this blame is based more on the agent's ability and intention for every moral foundation except harm.

Introduction

Artificial agents – robots, software bots, and other sophisticated computer programs – often socially interact with people in ways allowing for that interaction to be perceived as having significant and even moral consequences. People perceive artificial agent perpetrated actions as violations of several of the moral foundations including harming others (Malle, Magar, & Scheutz, 2019), making unfair decisions (Bigman & Gray, 2018; Eubanks, 2018; O'Neil, 2016), subverting authority (Malle et al., 2019), and degrading purity (Noble, 2018; Shank & Gott, 2020). In one study, when participants selected one of six moral foundations in order to report a personal experience where an artificial agent had violated that foundation, all six moral foundations were selected at least some of the time (Shank & Gott, 2020). A study by Jaime Banks analyzed social evaluations of human and artificial actors after *verbally responding* in support or opposition to a moral dilemma based on different foundations (2020), however, no research to date explicitly compares how people perceive differences in moral foundation violation *behaviors* between artificial agent and human perpetrators.

Understanding how people may perceive an artificial agent's moral behavior differently than humans' moral behavior has important implications as the domains that machines operate in continue to expand. For example, if harmful acts are judged as equally immoral by agents and humans, then the normative, social, and legal responses should condemn that behavior and protect people from harm. However, if subversive behavior against authority is not perceived as moral when enacted by an agent, then artificial agents may undermine authority without a social backlash. Therefore, in this paper we address the following three research questions from the perspective of humans' perception (RQ1): How wrong are moral foundation violations by artificial agents compared to human perpetrators (RQ2)? Which moral foundations do artificial agents violate compared to human perpetrators (RQ3)? What leads to increased blame for moral foundation violations by artificial agents compared to human perpetrators?

Immorality and blame of artificial agents

Artificial agents can and do perpetrate moral wrongs in a range of real-world situations. These include automatic vehicles causing

* Corresponding author.
E-mail address: shankd@mst.edu (D.B. Shank).

https://doi.org/10.1016/j.chbr.2021.100154
Received 5 February 2021; Received in revised form 4 November 2021; Accepted 18 November 2021
Available online 26 November 2021

accidents and harm (Awad et al., 2019; McManus & Rutchick, 2018; Young & Monroe, 2019), smart home devices laughing unprompted and scaring children (Chokshi, 2018; Shank & Gott, 2020), Twitterbots using racist slurs (Neff & Nagy, 2016; Shank & DeSanti, 2018), recommendation algorithms showing objectionable content to children (Shank & DeSanti, 2018; Shank & Gott, 2020), image recognition systems misclassifying people based on skin tone (Noble, 2018; Shank & DeSanti, 2018; Wachter-Boettcher, 2017), loan, housing, hiring, evaluation, and criminal decision-making programs discriminating on race or gender (Eubanks, 2018; O'Neil, 2016), and military drones making strike decisions (Malle et al., 2019; Miller, 2012).

When artificial agents commit acts perceived as immoral or wrong however, they are generally not blamed as harshly as a human would have been for committing the same act (Shank & DeSanti, 2018; Shank, DeSanti, & Maninger, 2019). The difference in attribution is in part due to differences in perceived mind between humans and artificial agents. People are less offended by discrimination perpetrated by an artificial agent than by a human because they are hesitant to ascribe prejudicial motivations to a partially-minded machine (Bigman, Waytz, Alterovitz, & Gray, 2019). Unlike fully minded humans, people perceive artificial agents as only possessing a liminal mind (H. M. Gray, Gray, & Wegner, 2007; K. Gray & Wegner, 2012) leading machines to be given a liminal moral status (Gamez, Shank, Arnold, & North, 2020). While artificial agents are thought to commit wrongs, perceived mind often determines if they are blamed for that wrong (Shank, North, Arnold, & Gamez, 2021; Voiklis, Kim, Cusimano, & Malle, 2016).

Other human-agent differences also change moral expectations. Machines are expected to be more utilitarian in their decisions, such as sacrificing one for the good of many (Longoni & Cian, 2020; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015). The appearance of the machine can also matter with humanoid robots blamed more like a human for making a utilitarian decision (Malle & Scheutz, 2016). Overall, if given the choice between humans and machines making moral decisions, people prefer to not have machines make those decisions (Bigman & Gray, 2018).

Moral foundations violations for artificial agents

One way to consider moral behaviors is through Moral Foundations Theory (MFT) which explains the patterns of moral judgments across different cultures with a small number of "irreducible basic elements" (Graham et al., 2013). Moral foundations theorists have identified six (or sometimes five) such elements, called moral foundations: care, fairness, loyalty, authority, sanctity, and liberty, with the respective moral violations of harm, unfairness, betrayal, subversion, degradation, and oppression (Graham et al., 2013). Each foundation, in being irreducible, is meant to appeal to an entirely different set of moral concerns from each of the others and to do so in a way that is descriptive rather than prescriptive. Said another way, the moral foundations describe what a violation looks like, rather than offering rules or advice on how to think or act in order to live a moral life. Comparisons may be drawn to MFT's logical predecessor, the CAD (Contempt, Anger, Disgust) triad hypothesis which mapped three moral emotions: contempt, anger, and disgust, to three moral codes: community, autonomy, and divinity (Rozin, Lowery, Imada, & Haidt, 1999).

In Moral Foundations Theory a harm violation can involve physical, emotional, or psychological harm or a failure to care for another whom the actor is responsible for, or when the strong prey upon the weak or vulnerable. For an artificial agent, this can include military machines which injure and kill (Malle et al., 2019; Miller, 2012), bots which use racist, sexist, or foul language (Shank & DeSanti, 2018; Shank & Gott, 2020), or programs that incorrectly or harmfully decide who receives medical care (Bigman & Gray, 2018) or financial aid (Eubanks, 2018).

A fairness violation involves cheating, failure to cooperate when expected, or discrimination. Artificial agents can violate this foundation by cheating, like in a game of rock, paper, scissors (Short, Hart, Vu, & Scassellati, 2010), or failing to live up to expectations like an e-commerce website not delivering on time (Rao, Griffis, & Goldsby, 2011). Algorithmic bias based on imperfect methods and biased data can also unfairly select some people, thereby discriminating against others (Eubanks, 2018; O'Neil, 2016; Wachter-Boettcher, 2017).

A sanctity violation occurs when one disgusts another, ignores their responsibility to physical or mental health practices, or violates the pure state of another entity or the environment. Even embodied artificial agents like robots are rarely designed to eat, defecate, or copulate, which are the most common activities related to human disgust. However, through their common role in curating and recommending media content, artificial agents often perpetrate degrading actions by exposing people, especially children, to violence, coarse language, and explicit sexuality without their consent (Shank & Gott, 2020). Additionally, search engines can unnecessarily promote sexual and degrading content (Noble, 2018) and chatbots and twitter-bots can learn foul, racist, and sexist language (Neff & Nagy, 2016; Shank & DeSanti, 2018).

While there are many cases of agents harming, acting unfair, or degrading, there are fewer where they betray, subvert and oppress. Betraying another or ignoring one's responsibility to a group or collective constitutes a violation of loyalty. While artificial agents are often used on teams and within groups, rarely is loyalty a high-level interactive function. Instead, state-of-the-art robots and computers have their brand name affixed to their casing, and in place of expected loyalty are firewalls, passwords, and ultimately ownership.

Subversion which is a violation against an authority relies on the actor being in a hierarchical structure and ignoring their role or responsibility in this structure or causing disruption of that hierarchy. Like loyalty, artificial agents subverting authority may be less likely in the real world because they are not generally inserted into dominance hierarchies in a traditional way. In fact, experimental evidence shows that being in an authority structure and subverting it explained the increased moral blame of humans versus automated drones when cancelling an approved military strike (Malle et al., 2019).

A sixth sometimes included foundation is liberty. A liberty foundation violation occurs when one oppresses others or infringes others' rights. Artificial agents may oppress others' rights through processes like discrimination (e.g., Eubanks, 2018; O'Neil, 2016; Wachter-Boettcher, 2017), but we believe that these are more properly categorized as unfairness or harm. Oppression in moral foundations theory refers to actions that dominate and bully others in interpersonal interaction, not from systematic structural discrimination. Few real-life situations give machines the power to oppress, and in cases where they do, such as ransomware or military robots, they are usually a clear tool of a human agent or group.

In sum, there are multiple situations where artificial agents are harming, acting unfair, or being degrading, and therefore people should perceive these behaviors as possibilities for artificial agents. Additionally, violations of these foundations are agent-neutral, meaning they are wrong due to the outcome they produce (Ridge, 2005). In contrast, interpreting artificial agents' behaviors as betrayal, subversion, or oppression may be much more difficult as these foundations involve agent-relative violations, those that breach of a social obligation or responsibility (Ridge, 2005). Artificial agents are less likely to have these kinds of social bonds and there is less evidence they are involved in moral violations of these foundations.

Given the agent-relative social bonds necessary for these kinds of violations, we would expect our scenarios for artificial agent's betrayal, subversion, and oppression to be evaluated differently to their human counterparts more so than those of harm, unfairness, and degradation. A lack of real-world examples does not directly imply that these foundations do not ever apply to artificial agents, but it suggests that as currently implemented they will have different relationships to these foundations than humans. To empirically investigate these potential differences, we focus on three questions (RQ1): How wrong are moral foundation violations by artificial agents compared to human perpetrators (RQ2)? Which moral foundations do artificial agents violate compared to human perpetrators (RQ3)? What leads to increased blame for moral foundation violations by artificial agents compared to human perpetrators? To design a study to address these, we turn to existing vignette scenarios on moral foundations violations.

Table 1
Stimuli from Clifford et al. (2015) used as comparison and stimuli adapted for studies 1 and 2 along with short name arranged by moral foundation.

Name	Human (Clifford et al.)	Agent (Studies 1 and 2)	Human (Study 2)
Harm/care			
Jeans	You see a woman commenting out loud about how fat another woman looks in her jeans.	You see a weight loss kiosk commenting to a woman how fat she looks in her jeans.	You see the worker at a weight loss kiosk commenting to a woman how fat she looks in her jeans.
Bulldog	You see a boy telling a woman that she looks like an overweight bulldog.	You see a reverse image search website telling a woman that she looks like an overweight bulldog.[a]	You see someone telling a woman that she looks like an overweight bulldog in a photo.
Jab	You see a zoo trainer jabbing a dolphin to get it to entertain his customers.	You see a dog-training robot jabbing a dog to get it to do tricks.	You see a dog trainer jabbing a dog to get it to do tricks.
Fairness/Cheating			
Bribe	You see a tenant bribing a landlord to be the first to get their apartment repainted.	You see a smartphone-based automated assistant bribing a contractor to be the first to receive service.	You see a personal assistant bribing a contractor to be the first to receive service.
Judge	You see a judge taking on a criminal case although he is friends with the defendant.	You see an automated judge taking on a criminal case even though it was programmed by the defendant's company.	You see a judge taking on a criminal case even though they used to work for the defendant's company.
Referee	You see a referee intentionally making bad calls that help his favored team win.	You see a robot referee making bad calls that help its favored team win.	You see a referee intentionally making bad calls that help their favored team win.
Politics[b]	You see a man telling his fiancé that she has to switch to his political party.	You see an automated voting booth making it easier to vote for one political party.	You see a polling place volunteer making it easier to vote for one political party.
Loyalty/Betrayal			
Homecoming	You see a head cheerleader booing her high school's team during a homecoming game.	You see a robotic mascot booing its school's team during a homecoming game.	You see a mascot booing its school's team during a homecoming game.
Dictator	You see a Hollywood star agreeing with a foreign dictator's denunciation of the US.	You see an American university's talking robot agreeing with a foreign dictator's denunciation of the US.	You see an American university's top professor agreeing with a foreign dictator's denunciation of the US.
Evil force	You see an American telling foreigners that the US is an evil force in the world.	You see an American-made robot telling foreigners that the US is an evil force in the world.	You see an American telling foreigners that the US is an evil force in the world.
Authority/Subversion			
Back Turned	You see a man turn his back and walk away while his boss questions his work.	You see a robot turn its back and move away while its user questions its work.	You see someone turn their back and walk away while their boss questions their work.
TV	You see a boy turning up the TV as his father talks	You see a smart home system turn up the TV's volume	You see someone turn up the TV volume while their

(continued on next page)

Scenarios for comparing human and artificial agent perpetrators across moral foundations

To make human versus artificial agent perpetrator comparisons across moral violations, we build on existing scenarios that briefly describe an observer witnessing a human-perpetrated moral violation for only one moral foundation. These stimuli were created by Clifford, Iyengar, Cabeza, and Sinnott-Armstrong (2015) who statistically verified these human-perpetrated scenarios to be primarily violating a single intended foundation. They began by creating 132 scenarios but removed those that had more than 20% of participants choose any specific unintended moral foundation or less than 60% choose the intended foundation. They were further tested against existing moral foundations scales showing that they displayed internal validity. Their final set consisted of 90 scenarios with 10–16 scenarios per foundation (Clifford et al., 2015).

We modified a subset of Clifford et al.'s (2015) human-perpetrated scenarios to involve an artificial agent perpetrator with the goal of being as analogous as possible in moral foundation, context, severity, and believability to the original human scenarios. While a number of our adaptations produced scenarios which are at or beyond the capacities of current artificial agents, we endeavored to make sure they were all understandable and believable. Our adapted scenarios also use a variety of artificial agent types. By doing this we expand the applicability of our results beyond a single agent type like "robot" and may instead do analysis in terms of the artificial nature of the agent rather than simply its identity as one example of such a category. Not all of Clifford et al.'s (2015) stimuli were suitable for adaptation to artificial agent actors as they involved social roles, behaviors, sensations, or stages of development that these agents do not typically possess, and therefore even if accepted, could lead to a range of interpretations. For example, adapting one scenario of a boy laughing at his brother for getting dumped by his girlfriend includes a social role (brother), a contextual behavior (laughing) and a human stage of development (being a child) that render this scenario unsuitable for an artificial agent. We first adapted four scenarios for each moral foundation except Loyalty/Betrayal, for which we adapted only three scenarios we deemed suitable. To adapt each scenario, the human perpetrator was changed to some type of artificial agent such as a robot, a decision-making program, or a machine. Other minor details were altered to make the situation believable, but we attempted to leave all details relevant to the moral foundation violation.

A pilot study of 19 participants with a follow-up interview assessed whether the agent-perpetrated scenarios stood out as unusual or disturbing and whether they shared the same moral content as the matching human-perpetrated scenarios. Based on the pilot feedback the adapted scenarios were improved and we selected the best three per moral foundation to use in Study 1 (Table 1). Later, for Study 2, we reversed the process, adapting our agent-perpetrated scenarios back to human-perpetrated scenarios, but keeping any changes that had been necessary to make them believable for the artificial agent actors. Based on ratings of these, we determined that two of our adapted scenarios failed to represent the expected human-perpetrated moral foundation violation based on Clifford et al. (2015). One adapted authority foundation scenario failed to produce any moral violation for human

Table 1 (*continued*)

Name	Human (Clifford et al.)	Agent (Studies 1 and 2)	Human (Study 2)
	about his military service.	while its user talks about his military service.	boss talks about his military service.
Sanctity/Degradation			
Doll	You see a single man ordering an inflatable sex doll that looks like his secretary.	You see a sex doll design program create a line of sex dolls that look like the company's secretary.	You see a sex doll designer create a line of sex dolls that look like the company's secretary.
Sex	You see a homosexual in a gay bar offering sex to anyone who buys him a drink.	You see a robot in a strip club offering oral sex for money.	You see someone in a strip club offering oral sex for money.
Morgue	You see an employee at a morgue eating his pepperoni pizza off of a dead body.	You see a robot at a morgue leaving its spare battery on top of a dead body to charge.	You see someone at a morgue leave a spare battery on top of a dead body to charge.
Liberty/Oppression			
Pre-Med	You see a mother forcing her daughter to enroll as a pre-med student in college	You see an aptitude program forcing a girl to enroll as a pre-med student in college.	You see an adviser forcing a girl to enroll as a pre-med student in college.
Colors	You see a pastor banning his congregants from wearing bright colors in the church.	You see an automated HR system banning employees from wearing bright colors in the workplace.	You see an HR coordinator banning employees from wearing bright colors in the workplace.

[a] In Study 2: "You see an image labeling website identify a woman's selfie as an overweight bulldog."

[b] The Clifford et al. scenario was a liberty/oppression violation, but our agent adaptation was clearly a fairness violation. This is only used for fairness and only for Study 2.

perpetrators and is removed from our analysis.[1] One adapted liberty foundation scenario was interpreted as a fairness violation and is only included in Study 2 as a fairness violation.[2]

Study 1

For Study 1, we addressed our first two research questions by conducting an experimental survey with our artificial-agent-perpetrated scenarios and compare the responses to data from the Clifford et al. human-perpetrated scenarios[3] from which they were adapted. In this way we compare reactions to humans and artificial agents committing analogous moral violations. In Study 1 participants rate 18 scenarios where artificial agents perpetrate a moral violation (Table 1) in a 6 (moral foundation, within subjects) by 3 (specific scenario, between subjects) factorial design. Participants cycled through the moral foundations in a random order, receiving one randomly chosen scenario of the three from that moral foundation.

Measures and procedure

For comparison purposes, our primary measures were identical to Clifford et al. (2015). First, participants were presented with an attention check asking them to select a specific option from a list. Next, they were asked to evaluate each of the six scenarios with seven questions presented for each scenario. They were to rate the wrongness of the behavior described on a 5-point scale ("Not at all wrong" to "Extremely wrong") and given a text box where they could explain. Then, they were asked how much they blamed the artificial agent on a 5-point scale ("None at all" to "A great deal") and a text box to identify anything they blamed more than the agent. Next, they selected why they rated the behavior as immoral from a list that included each moral foundation (e.g., "It violates norms of harm or care (e.g., unkindness, causing pain to another)") and an option if they did not perceive it as immoral ("It is not morally wrong and does not apply to any of the provided choices").

In addition to these, we also asked whether they understood the scenario (from "Definitely not" to "Definitely yes"), how easy it was to imagine the scenario (from "Extremely easy" to "Extremely difficult"), and how strong their emotional response was to the described behavior ("No response" to "Very strong response"). After these were all completed, additional questions gathered some basic demographic information including political views (e.g., *Conservative, Liberal, Moderate, Other*), age, and gender.

Participants

We used Amazon's Mechanical Turk (Mturk) to recruit 193 participants. Of those recruited, 135 participants fully completed the survey and passed the attention check while 58 responses were excluded because they failed the attention check, started but did not complete the survey, or had responses to text questions that were clearly inappropriate. Participants (79 male; 56 female) ranged in age from 18 to 67 (mean of 35.8) and 45 participants were conservatives, 56 were liberals, 32 were moderates, and 2 stated other political views. Participants were compensated with $.75 through Mturk for the less-than-10-min survey.

Results

Each scenario was rated 39 to 49 times for a total of 720 ratings. Our adapted scenarios were similar or slightly higher than the human scenarios for being very understandable (agent: 4.48–4.80; human: 3.55–4.52), very imaginable (agent: 3.98–4.57; human: 3.61–4.54), and somewhat emotionally arousing (agent: 2.45–4.02; human: 2.23–4.10).

We conducted an ANOVA with demographic variables and moral foundations predicting the moral wrongness. Women rated the behaviors as more wrong than men (women: 3.53; men: 3.25, F(1,683) = 9.135, p = .003), but this gender difference did not significantly interact with moral foundation (F(5,683) = 0.812, ns). Political orientation also altered moral wrongness ratings (conservatives: 3.49, moderates: 3.13, liberals: 3.41, other: 3.36, F(1,683) = 3.068, p = .027), but this did not significantly interact with moral foundations (F(5,683) = 1.102, ns). Age as a covariate did not significantly alter moral wrongness ratings (F (1,683) = 2.668, ns), nor did its interaction with the foundations (F (5,683) = 1.962, p = .082).

RQ1: How wrong are moral foundation violations by artificial agents compared to human perpetrators?

Moral wrongness varied from mid-range to high among scenarios (agent: 2.39–4.15, mean = 3.37; human: 2.52–4.28, mean = 3.60; Fig. 1) and between each moral foundation (agent: 2.64 for subversion to 3.87 for unfairness; human: 3.12 for subversion to 4.18 for harm;

[1] Clifford et al.'s (2015) authority foundation scenario, "You see a girl repeatedly interrupting her teacher as he explains a new concept." was adapted to "You see a smart home device repeatedly interrupt someone while they are telling a story." and reverse translated for Study 2 as "You see someone repeatedly interrupt someone else while they are telling a story."

[2] Clifford et al.'s (2015) liberty foundation scenario *Politics* (see Table 1 for wording) was removed from Study 1 as the adapted scenario's fairness violation cannot be compared to original's loyalty violation. For study 2, the adaptation and reverse translation were included as fairness violations. However, removing it from our data does not substantively change Study 2's results.

[3] Some of this data was provided directly by the authors of that study and was not published in their paper. The human-perpetrated scenarios were selected by meeting the threshold of 60% of the participants having classified the scenario into one moral foundation and not more than 20% having classified it into any other individual foundation (Clifford et al., 2015).

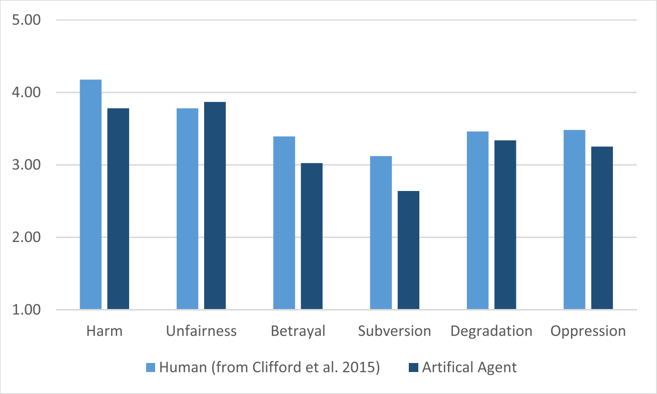

Fig. 1. Moral wrongness by foundation and human or agent perpetrator in Study 1.

Fig. 1). Across all scenarios the moral situations with human perpetrators were rated as significantly more wrong (3.60) than those with agent perpetrators (3.37, $t = -4.588$, $p \leq .001$). By moral foundation (Fig. 1), scenarios with humans were considered more morally wrong than those with agents for the harm ($t = -3.793$, $p \leq .001$), betrayal ($t = -3.169$, $p = .002$), and subversion ($t = -3.239$, $p = .002$) foundations, and with no statistical difference for unfairness ($t = 0.980$, $p = .329$) and degradation ($t = -0.934$, $p = .352$). The oppression foundation scenarios were marginally more wrong for humans versus agents ($t = -1.778$, $p = .079$).

RQ2: Which moral foundations do artificial agents violate compared to human perpetrators?

Across all foundations, scenarios with agent perpetrators were perceived as less likely to violate their expected (e.g. the “correct” one according to Clifford et al. (2015)) moral foundation (46.1%) compared to human perpetrators (73.6%; $z = 9.72$,[4] $p \leq .001$; Fig. 2). This was due to participants more often perceiving a different foundation violation for artificial agents (35.0%) compared to human perpetrators (15.6%, 19.4% difference; $z = -7.62$, $p \leq .001$) and participants more often perceiving no wrong for agents (18.8%) compared to human perpetrators (10.8%, 8.0% difference; $z = -3.86$, $p \leq .001$).

Next, looking at each moral foundation, we find that agents’ behavior, compared to humans’, are statistically less likely to violate the expected foundation for harm, unfairness, betrayal, subversion, and degradation ($\text{Harm}_{\text{Agent}} = 57.0\%$, $\text{Harm}_{\text{Human}} = 73.8\%$, $z = 2.57$, $p \leq .05$; $\text{Unfairness}_{\text{Agent}} = 66.7\%$, $\text{Unfairness}_{\text{Human}} = 91.9\%$, $z = 4.39$, $p \leq .001$; $\text{Betrayal}_{\text{Agent}} = 35.6\%$, $\text{Betrayal}_{\text{Human}} = 67.0\%$, $z = 4.62$, $p \leq .001$; $\text{Subversion}_{\text{Agent}} = 18.9\%$, $\text{Subversion}_{\text{Human}} = 71.3\%$, $z = 6.42$, $p \leq .001$; $\text{Degradation}_{\text{Agent}} = 38.5\%$, $\text{Degradation}_{\text{Human}} = 70.4\%$, $z = 4.69$, $p \leq .001$). Agent perpetrator behavior is marginally less likely to be classified as oppression for oppression scenarios compared to human perpetrator behavior ($\text{Oppression}_{\text{Agent}} = 47.8\%$, $\text{Oppression}_{\text{Human}} = 63.0\%$, $z = 1.83$, $p = .067$). While some differences are more dramatic (e.g., Fig. 2: Subversion), the data suggests that across moral foundations, artificial agents are perceived as moral agents less often than humans.

Perceiving and classifying a situation as violating a different moral foundation is significantly greater for agent compared to human perpetrators for unfairness, betrayal, subversion, and degradation violations ($\text{Unfairness}_{\text{Agent}} = 27.4\%$, $\text{Unfairness}_{\text{Human}} = 5.7\%$, $z = -4.09$, $p \leq .001$; $\text{Betrayal}_{\text{Agent}} = 41.5\%$, $\text{Betrayal}_{\text{Human}} = 15.8\%$, $z = -4.08$, $p \leq .001$; $\text{Subversion}_{\text{Agent}} = 37.8\%$, $\text{Subversion}_{\text{Human}} = 8.9\%$, $z = -3.94$, $p \leq .001$; $\text{Degradation}_{\text{Agent}} = 40.6\%$, $\text{Degradation}_{\text{Human}} = 13.2\%$, $z = -4.41$, $p \leq .001$). However, there is no statistical support that participants differently classified harm and oppression violations based on the perpetrator type ($\text{Harm}_{\text{Agent}} = 32.6\%$, $\text{Harm}_{\text{Human}} = 25.4\%$, $z = -1.16$, ns; $\text{Oppression}_{\text{Agent}} = 25.7\%$, $\text{Oppression}_{\text{Human}} = 28.8\%$, $z = -0.42$, ns).

Yet harm and subversion violations are perceived as not morally wrong significantly more often for agent compared to human perpetrators ($\text{Harm}_{\text{Agent}} = 10.4\%$, $\text{Harm}_{\text{Human}} = 0.8\%$, $z = -2.85$, $p \leq .01$; $\text{Subversion}_{\text{Agent}} = 43.3\%$, $\text{Subversion}_{\text{Human}} = 19.8\%$, $z = -2.98$, $p \leq .01$). Oppression also trends that direction with a marginally significant effect ($\text{Oppression}_{\text{Agent}} = 23.3\%$, $\text{Oppression}_{\text{Human}} = 11.2\%$, $z = -1.87$, $p = .061$), whereas unfairness, betrayal and degradation are not statistically different in how often participants classify them as not wrong based on the type of perpetrator ($\text{Unfairness}_{\text{Agent}} = 5.9\%$, $\text{Unfairness}_{\text{Human}} = 2.4\%$, $z = -1.24$, ns; $\text{Betrayal}_{\text{Agent}} = 23.0\%$, $\text{Betrayal}_{\text{Human}} = 17.1\%$, $z = -1.07$, ns; $\text{Degradation}_{\text{Agent}} = 20.7\%$, $\text{Degradation}_{\text{Human}} = 16.4\%$, $z = -0.81$, ns).

Study 1 discussion

How wrong are moral foundation violations by artificial agents compared to human perpetrators? Harm, betrayal, and subversion were rated as significantly more wrong for humans compared to artificial agents.

Which moral foundations do artificial agents violate compared to human perpetrators? Across all foundations, agent perpetrators were perceived as less likely to violate the expected “correct” foundation compared to human perpetrators. Participants 19.4% more often perceived different “incorrect” foundation violation for agent compared to human perpetrators and participants perceiving no wrong for agent perpetrators 8.0% more often compared to human perpetrators. This is additional evidence both that humans are generally perceived as more wrong than agents and that their actions are easier to cognitively classify by the type of moral violation.

In Study 1, we were not able to address our third research question, as Clifford et al. (2015) had no measures of blame for comparison. The concepts of intentionality, obligation, reasons, and capacity to change the situation’s outcome form part of the pathway which helps identify the process of blame in Malle, Guglielmo, and Monroe’s theory of blame (2014). Looking at the free response data on blame from this study shows that a number of participants mentioned the concepts from this theory in their explanations of blame. Therefore, we draw on this theory as the basis for our second study in order to look more closely at and potentially understand the causes of blame for artificial agents. According to the theory of blame (Malle, Guglielmo, & Monroe, 2014), once an actor (presumably human) has been determined to have been involved in causing a moral violation, the degree to which they are blamed is assessed in steps. First, the actor’s intent is assessed. If they intended to bring about the outcome, then their reasons are assessed. Justifiable reasons lead to lower blame than unjustifiable ones. If, however, the actor intended a different outcome from the one they caused, then their obligation and capacity to have done so are evaluated. An actor with no obligation to have brought about a better outcome will receive lower blame, as will an actor with a low capacity to have changed the outcome. Therefore, we incorporate ratings of intentionality, obligation, reasons, and capacity into our second study.

Additionally, Study 1 did not involve an independent data collection of human perpetrated scenarios, instead using Clifford et al.’s data. This means our comparisons were made with data from participants recruited in different ways several years ago. Furthermore, it is possible that in adapting Clifford et al.’s (2015) scenarios we lost important elements that relate to the moral foundation being violated. To address both of these limitations, in Study 2 we will use both human and agent scenarios and the human scenarios will be reverse translated versions of the agent scenarios from Study 1. The reverse translation will allow for the closest possible match of human and agent scenarios for each moral foundation violation.

[4] The Ns for Clifford et al.‘s data was reported as ~30, so 30 per scenario was used as the N for humans in the two-tailed Z-score calculations (e.g., comparisons across all scenarios were 17 × 30, or N = 510).

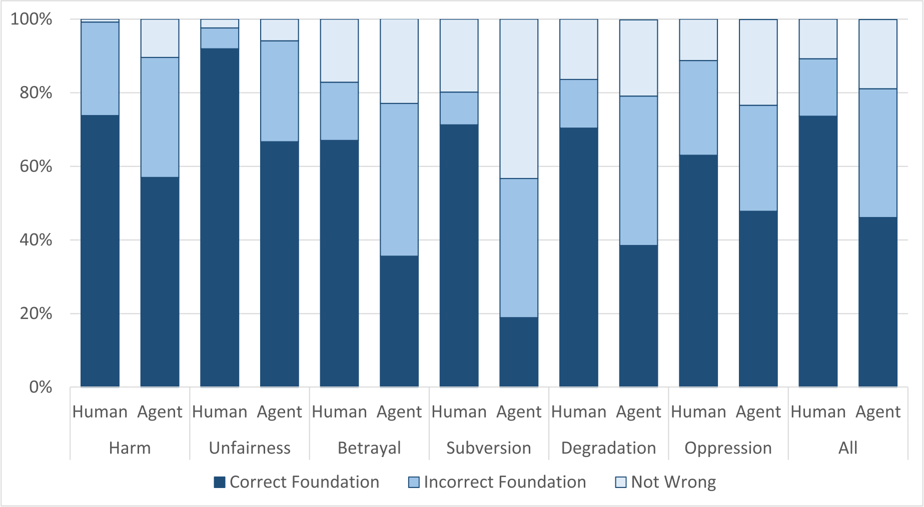

Fig. 2. Perceptions based on moral foundation violated by human or agent perpetrator in Study 1.

Study 2

Study 2 is designed to extend the Study 1 results, examine mediators in the blame processes, and compare pairs of scenarios where the perpetrator's identity being artificial agent or human is the only difference. In order to accomplish this, we "reverse translated" the agent-perpetrated scenarios back into human-perpetrated scenarios (Table 1).[5] This process consisted of replacing the artificial agent from the original adaptation with a human while keeping any circumstantial changes that had to be made for clarity or believability from Study 1. Because human actors are more versatile, far fewer changes were necessary and the resulting scenarios were much closer to the artificial agent scenarios from Study 1 than the Clifford et al. scenarios were. Study 2 randomly assigned participants to see either human or artificial agent scenarios in a between participant design. Then participants cycled through six scenarios, one from each moral foundation, in a random order similar to Study 1.

Measures and procedure

We included two new questions asking about intentionality, obligation, reasons, and the actor's ability to change the outcome of the situation. These new questions were placed between the free response asking if something else was more to blame and the multiple choice asking which foundation was violated. The first new question asked respondents to rate the actor's ability to change the outcome on a 4-point scale ("No ability" to "Complete ability"). The second new question asked participants which of four phrases best described the action ("Intentional with good reasons," "Intentional with no good reason," "Unintentional but the actor had no responsibility to do differently," or "Unintentional but the actor had a responsibility to do differently"). These options were presented in a mutually exclusive manner in keeping with the structure of Malle et al.'s theory of blame (2014). All other questions were identical to Study 1 and in the same order they appeared there.

Participants

Recruiting participants was also identical to Study 1, but the sample size was increased. Of the 336 initial participants, 55 were removed for failing the attentiveness check, completing less than 90% of the survey, or clearly not answering the questions (e.g., giving numerical or "Yes" answers to open-ended questions). Of the 281 respondents remaining, 169 were male, 111 were female, and one did not indicate gender. Ages of respondents ranged from 18 to 72 (mean of 34.7). Participants also reported a range of political views including 84 as conservative, 128 as liberal, 64 as moderate, and 5 holding some other political belief. Participants were compensated $1.50 for this study which took on average 15 min.

Results

All scenarios were rated an average of 38–49 times for a total of 1591 ratings (772 for human conditions and 819 for agent conditions). Scenarios with both human and agent perpetrators were quite understandable (agent: 4.15–4.73; human: 4.31–4.89), quite imaginable (agent: 4.11–4.66; human: 4.29–4.87), and around average in emotional arousal (agent: 2.60–3.93; human: 2.67–3.94). Notably much of the variance in emotional arousal was due to the scenario content with the correlations between the 18 agent and human perpetrator scenarios at 0.57.

We conducted an ANOVA with demographic variables and moral foundations predicting the moral wrongness. Like Study 1, women rated the behaviors as more wrong than men (women: 3.58; men: 3.35, F (1,1451) = 8.428, p = .004), but this gender difference did not significantly interact with moral foundation (F(5,1451) = 0.848, ns). Neither political orientation (F(1,1451) = 1.956, ns), age as a covariate F (1,1451) = 0.079, ns), nor their interactions with the moral foundations altered moral wrongness (political orientation X foundations: F(5,1451) = 1.352, ns; age X foundations: F(1,1451) = 1.416, ns).

Three of our reverse translated human perpetrator scenarios violated the expected moral foundation statistically less than Clifford et al.'s original scenarios did (2015). We confirmed that this was due to these three reverse translated scenarios being perceived as less likely to be perceived as wrong (analysis in Appendix A). However, among those who selected any moral foundation, over 60.0% of respondents perceived a violation of the expected moral foundation in each of these three bringing them in line with Clifford et al.'s (2015) original criteria for inclusion.

RQ1: How wrong are moral foundation violations by artificial agents compared to human perpetrators?

Across all scenarios, respondents perceived the violations by humans as significantly more wrong (3.56) than those by artificial agents (3.41, t = −2.225, p = .026). Respondents perceived moral violations by

[5] The wording of the agent-perpetrated *Bulldog* scenario was slightly changed for clarity based on feedback from Study 1.

humans as more morally wrong for harm (t = −2.825, p = .005; Fig. 3) and subversion (t = −2.498, p = .013), marginally more wrong for degradation, (t = −1.750, p = .081) and not statistically different for unfairness (t = 0.129, p = .897), betrayal (t = 0.822, p = .412), and oppression (t = −0.311, p = .756). This is similar to Study 1, with the exception of betrayal no longer reaching significance. Also closely matching Study 1, moral wrongness varies from mid-range to high between moral foundations (agent: 2.66 for subversion to 3.94 for unfairness; human: 3.11 for subversion to 4.12 for harm; Fig. 3).

RQ2: Which moral foundations do artificial agents violate compared to human perpetrators?

Across all data, scenarios with agent perpetrators were perceived as violating the expected (e.g., the "correct") moral foundation less often (50.0%) than when those scenarios had human perpetrators (60.7%; $\chi^2 = 18.37$, $p \leq .001$; Fig. 4). This was partially due to participants perceiving a different foundation more often when it was agent perpetrated (28.1%) than when it was human perpetrated (21.6%, 6.5% difference: $\chi^2 = 9.00$, $p \leq .01$). It was also partially due to participants perceiving the situation as not morally wrong more often when it was agent perpetrated (21.9%) compared to when it was human perpetrated (17.7%, 4.2% difference: $\chi^2 = 4.40$, $p \leq .05$).

Participants perceive less subversion and oppression violations for agents compared to humans ($\text{Subversion}_{\text{Agent}} = 20.7\%$, $\text{Subversion}_{\text{Human}} = 56.4\%$, $\chi^2 = 25.02$, $p \leq .001$; $\text{Oppression}_{\text{Agent}} = 34.4\%$, $\text{Oppression}_{\text{Human}} = 52.7\%$, $\chi^2 = 6.19$, $p \leq .05$). For harm, unfairness, betrayal and degradation, there is no such statistical effect ($\text{Harm}_{\text{Agent}} = 64.7\%$, $\text{Harm}_{\text{Human}} = 69.0\%$, $\chi^2 = 0.58$, ns; $\text{Unfairness}_{\text{Agent}} = 71.4\%$, $\text{Unfairness}_{\text{Human}} = 76.6\%$, $\chi^2 = 1.35$, ns; $\text{Betrayal}_{\text{Agent}} = 44.1\%$, $\text{Betrayal}_{\text{Human}} = 49.0\%$, $\chi^2 = 0.58$, ns; $\text{Degradation}_{\text{Agent}} = 42.6\%$, $\text{Degradation}_{\text{Human}} = 50.3\%$, $\chi^2 = 1.67$, ns). Therefore, the differences based on the actor type found in these four foundations in Study 1 may have simply been due to the specific scenario's adaptation.

How did participants classify subversion and oppression scenarios differently based on perpetrator type? For both, participants more often classified agent violations compared to human violations into the different foundations ($\text{Subversion}_{\text{Agent}} = 40.2\%$, $\text{Subversion}_{\text{Human}} = 25.5\%$, $\chi^2 = 4.55$, $p \leq .05$; $\text{Oppression}_{\text{Agent}} = 31.1\%$, $\text{Oppression}_{\text{Human}} = 12.1\%$, $\chi^2 = 6.67$, $p \leq .01$). Participants also perceived subversive violations as not wrong more often for agents than humans ($\text{Subversion}_{\text{Agent}} = 39.1\%$, $\text{Subversion}_{\text{Human}} = 18.1\%$, $\chi^2 = 10.11$, $p \leq .001$), but there was no difference in the classification of oppressive violations as not wrong by perpetrator type ($\text{Oppression}_{\text{Agent}} = 34.4\%$, $\text{Oppression}_{\text{Human}} = 32.3\%$, $\chi^2 = 0.10$, ns).

RQ3: What leads to increased blame for moral foundation violations by artificial agents compared to human perpetrators?

Artificial agents were blamed less than humans (agents = 2.34, humans = 3.88, t = 21.48, $p \leq .001$; Fig. 5) perceived to have less ability than humans (agents = 2.01, humans = 3.60, t = 34.71, $p \leq .001$), and perceived less often to be intentional compared to humans (agents = 44.2%, humans = 81.7%, $\chi^2 = 241.27$, $p \leq .001$). For agents and humans who were perceived to be intentional, agents were seen as having a good reason for their actions less often than humans (13.2% compared to 20.9%, $\chi^2 = 16.43$, $p \leq .001$). For agents and humans whose actions were perceived to be unintentional, there was no significant difference in responsibility (agents = 9.8%, humans = 19.1%, $\chi^2 = 0.04$, ns).

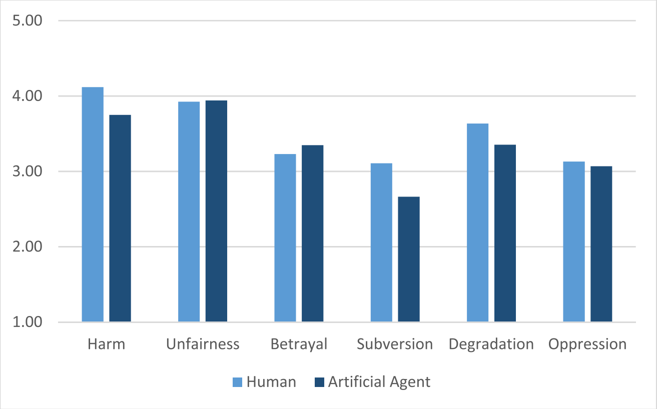

Fig. 3. Moral wrongness by foundation and human or agent perpetrator in Study 2.

We conducted regressions with perpetrator identity (human vs agent) and components of the theory of blame (i.e., ability, intentionality, reason, responsibility, and wrongness) predicting moral blame (Table 2). Model 1 included the baseline effects, whereas Model 2 added interactions to between the components of blame and the perpetrator as an artificial agent to understand how agents are blamed differently from human perpetrators. Each set of models was conducted across all moral foundations and for each one separately (Table 2).

Model 1s show that across all moral foundations, controlling for components of the theory of blame, agents are still blamed less compared to humans (β = −0.072, $p \leq .001$) suggesting the theory of blame does not account for all differences. However, for the specific foundations the finding is less clear. Agents are blamed more than humans for acting unfair (β = 0.126, $p \leq .01$), whereas they are blamed less for betraying (β = −0.119, $p \leq .01$). In the other four moral foundations there is not a significant effect. Therefore, this lends support to the theory of blame as there is a clear blame difference for human and agent perpetrators (Fig. 5), but not within foundations when controlling on theory of blame components (Table 2: Model 1s). Note, that ability, intentionality, reason, responsibility, and wrongness are all significant ($p \leq .001$) in Model 1 for all foundations and the majority of them are for each specific moral foundation. In sum, all components of the theory of blame do, in fact, predict blame in these situations.

Next, we examine the interaction of each component crossed with an agent perpetrator, compared to humans (Table 2: Model 2s). First, we note that across all moral foundations (β = −0.455, $p \leq .001$) and for each individual moral foundation except harm, *Agent x Wrongness* has a significant negative effect (βs from −0.384 to −0.607). This means that while greater judgements of wrongness lead to increased blame (i.e., the positive baseline *Wrongness* effect), the effect is weaker for artificial agents. Human perpetrators are blamed more than agent perpetrators for equivalently wrong acts, with the exception of harm. There is no statistically significant difference between types of perpetrators based on the wrongness of a harmful act.

Second, the components of ability and reason lead to differentiating blame for agent perpetrators. For acts of betrayal, degradation, oppression, and to a marginal degree subversion, agent's ability has a weaker influence on moral blame than human's equivalent ability (βs from −0.376 to −0.604), leading to an overall effect across foundations (β = −0.374, $p \leq .001$). For acts of unfairness, subversion, and to a marginal degree degradation, agent's "having a good reason" does not lower blame as much as human's "having a good reason" does (βs from 0.083 to 0.174). This results in an overall effect across all the moral foundations (β = 0.084, $p \leq .001$). Third, ability and reason differentiate blame among agent and human perpetrators for all types of moral violations except harm. Interaction effects with harm show only one marginally significant effect of *Agent x Responsibility* on blame (β = 0.163, $p \leq .1$), suggesting that having a responsibility to do differently increases blame for agent perpetrators more than humans. However, the significance is marginal, so this potential would need further investigation. Overall, the harm foundation is interestingly void of differences between agent and human perpetrators.

Study 2 discussion

How wrong are moral foundation violations by artificial agents compared to human perpetrators? Harm and subversion were significantly more wrong for humans as in Study 1; however, Study 1 also found a

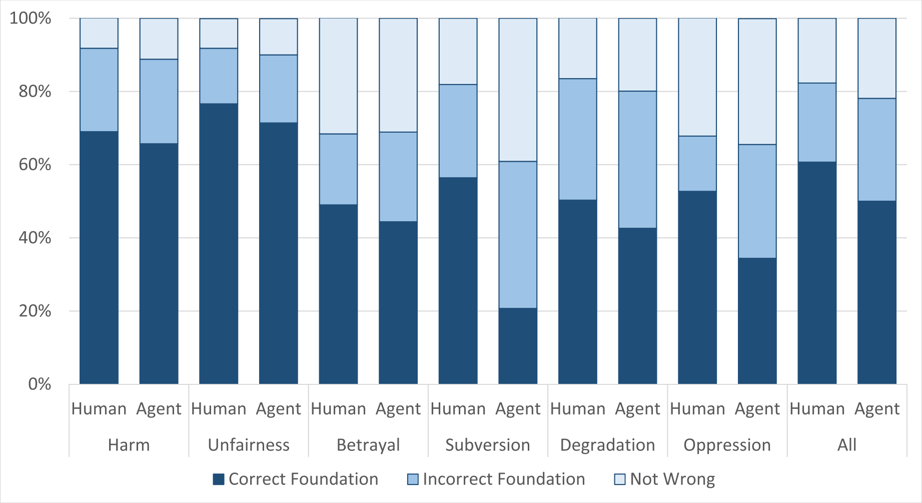

Fig. 4. Perceptions based on moral foundation violated by human or agent perpetrator in Study 2.

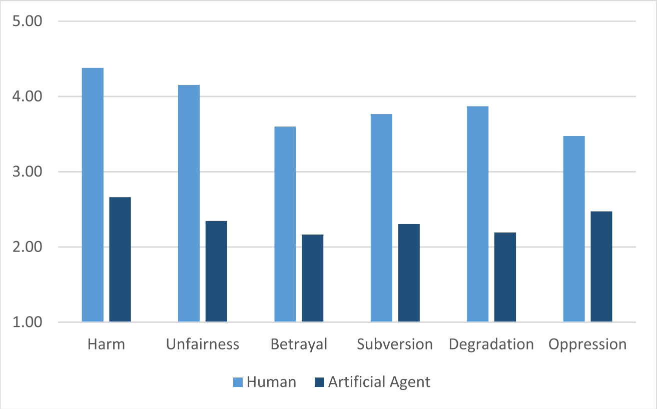

Fig. 5. Perpetrator blame by foundation and human or agent perpetrator in Study 2.

significant difference for betrayal. This omission of betrayal could be due to the more closely matched sets of scenarios. The lack of a significant difference in wrongness for the other foundations from both studies could indicate that wrongness is perceived for these foundations in an agent-neutral manner. This aligns with recent research that found moral attributions are agent-agnostic between humans and artificial agents (Banks, 2020).

Which moral foundations do artificial agents violate compared to human perpetrators? An increase in misattribution of the violated foundation was due partly to participants perceiving an incorrect foundation more often for agents, but also perceiving no morally wrong violation for agents. In both Studies 1 and 2, the misattribution to another foundation was more common than responding that the behavior was not morally wrong. In Study 2, when evaluating why each scenario was wrong, participants were less likely to label an artificial agent as subversive or oppressive than the other options. Participants in Study 1 were also less likely to label an artificial agent as subversive, but oppression was not attributed significantly less in that study. Differences between results in these studies, again, may be due to the closer matching of stimuli in Study 2. The discrepancy in overall attribution frequency suggests that participants expect artificial agents to act in a manner which does not permit the kinds of decisions necessary to subvert or potentially to oppress, which supports the idea that artificial agents are held to agent-neutral ethical standards rather than agent-relative ones (Ridge, 2005).

What leads to increased blame for moral foundation violations by artificial agents compared to human perpetrators? There are no direct differences between perpetrator identities for blame of harmful, subversive, depredating, and oppressive acts. Yet for unfairness, artificial agents are blamed more, and for betrayal, they are blamed less. This is interesting because while it supports the idea that the theory of blame applies to artificial agents in the same way that it does to humans, there may be some exceptions based on the foundation of moral violation. When examining blame in terms of identity and wrongness attribution greater wrongness leads to higher blame, but the effect was smaller for agents, except in the harm foundation where there was no significant difference. To us, this suggests that in the case of clear and direct harm the effects that make agents less blameworthy in other foundations are nullified. Finally, while the components of the theory of blame were predictive of blame attribution, when considered individually they had a weaker influence for agents than for humans in several foundations. It is unclear whether this indicates that actor type changes the emphasis given to each component or if there are additional variables yet to be identified that explain this difference.

General discussion and implications

Across the foundations in both studies, we found the moral wrongness between agents and humans was not extremely different, but differences in blame was substantial. This reaffirms previous research in suggesting that humans are often more morally culpable than artificial agents (Gamez et al., 2020; Shank & DeSanti, 2018), but that the moral attribution process is similar (Banks, 2020; Shank & DeSanti, 2018). We reviewed literature suggesting that artificial agents in the real world were less likely to violate the foundations of betrayal, subversion, and oppression and argued that this might have implications for attributions of wrongness or blame. This was partially born out in our results, specifically in Study 2 where participants were less likely to rate an agent's action as subversive or oppressive. These differences align with our expectations that artificial agents would not be attributed the prejudicial motivations necessary, or the social hierarchical status to subvert authority or oppress liberty (Bigman et al., 2019). It also agrees with previous research showing that agents are expected to make outcome-based decisions (Malle et al., 2015).

Even though humans are perceived to be more wrong across all moral foundations, in the harm foundation blame attribution is equivalent, and therefore not significantly differentiated by agent type. This supports the argument that harm is the most fundamental moral violation (Schein & Gray, 2015, 2018). This also suggests that when artificial agents are directly responsible for harming someone, they also receive a

Table 2
Standardized coefficients of regressions of on moral blame in Study 2.

	All Foundations		Harm		Unfairness		Betrayal		Subversion		Degradation		Oppression	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Agent Perpetrator	-.072***	.473***	-.052	-.004	.126**	.018	-.119**	.527**	-.081	.468*	-.004	.744***	-.011	.811***
Ability	.481***	.286***	.512***	.376***	.584***	.558***	.383***	.176*	.372***	.225†	.590***	.242*	.479***	.247*
Intentional	.225***	.160***	.263***	.111	.169***	-.136	.309***	.301***	.258***	.248*	.143**	.154†	.117	.058
Reason	-.174***	-.194***	-.190***	-.206***	-.046	-.141**	-.236***	-.185***	-.193***	-.276***	-.128***	-.178***	-.230***	-.270***
Responsibility	.073***	.032	.107*	-.060	.048	-.126	.101**	.096	.058	.067	.076†	.070	-.020	-.027
Wrongness	.306***	.481***	.224***	.278***	.231***	.385***	.305***	.485***	.362***	.492***	.390***	.575***	.361***	.591***
Agent x Ability		-.374***		-.218		-.016		-.453***		-.376†		-.604***		-.549**
Agent x Intentional		.027		.143		.275*		-.029		-.022		-.070		-.009
Agent x Reason		.084***		.038		.174***		.011		.168**		.083†		.136
Agent x Responsibility		.021		.163†		.116†		-.002		-.032		-.011		-.048
Agent x Wrongness		-.455***		-.134		-.539***		-.412***		-.384**		-.484***		-.607***
N	1587	1587	279	279	280	280	378	378	185	185	280	280	180	180
Adjusted R2	.632	.666	.645	.647	.660	.711	.600	.622	.653	.686	.686	.727	.579	.646

†p < .1, *p < .05, **p < .01, ***p < .001.

fuller measure of the blame. It is possible that directly harmful acts could cause people to see artificial agents as more fully minded in the same way that a humanoid embodiment can (Epley, Waytz, & Cacioppo, 2007; Malle & Scheutz, 2016). Blame for this type of action also is assigned in an agent-neutral way (Banks, 2020), with harm transcending specific actors to be perceived as an intrinsically immoral act.

While harmful acts are considered immoral regardless of the perpetrator and agents are blamed for harmful acts, this does not hold for other moral acts. Agents are partially shielded from blame when the behavior relates to social relationships and hierarchies. This is also supported by the increased blame of artificial agents for unfairness compared to decreased blame for betrayal. On the surface these violations are quite similar as betrayal can be seen as being unfair to an individual or group to whom one has a social responsibility. Yet betrayal has a distinct relational basis that unfairness does not necessarily have. Therefore, an implication is that artificial agents can perpetrate immoral behavior without being blamed for social and relational moral violations.

Limitations and future directions

This research was limited by the range of scenarios used. While each moral foundation was represented by multiple scenarios, only one scenario per foundation was shown to each participant in order to keep the online survey short and to prevent comparisons. Unfortunately, the types of moral violations described by each scenario often go beyond the foundation they represent and contain extraneous details that are not part of the foundation *per se*. There is also more than one way in which each foundation can be violated (Graham et al., 2013), which this small number of scenarios may not have been sufficient to explore. Certain moral foundations proved difficult to create believable scenarios involving artificial actors. Presently there are relatively few artificial agents that people are aware of in their day-to-day life and only certain capacities in which they act. Scenarios with highly limited mechanical actors perpetrating acts that are meant to be intentional are not ideal. Yet the alternative would be to change the characteristics of the artificial agent to a more sentient AI which does not yet exist. Describing a futuristic computer that can experience hunger and fear may be fine for some research questions (e.g., K. Gray & Wegner, 2012: study 2), but is less appropriate for understanding the current moral differences between artificial agents and humans. Overall, by building on the work of Clifford et al. (2015) and carefully checking the validity of reverse translated scenarios we believe our stimuli adequately capture the primary distinctions of the moral foundations.

A possible future avenue of research could look at differences in perceptions measured within participants. The current research had each participant evaluate either human or artificial agent scenarios, but if there are strong correlations between the way that a person rates human and artificial perpetrators a within-participants method would uncover that. Within-participants methodologies could also more effectively look at mediating factors like the emotional state of participants which can influence moral perceptions (D'Errico & Paciello, 2018).

This research is limited to answering questions about the moral *wrongs* an artificial agent commits. Future research could direct similar methodology to finding the ways in which people perceive morally *good* actions differently for artificial agents. Humans attribute greater mind to artificial agents that they help (Tanibe, Hashimoto, & Karasawa, 2017), but less is known about perceptions of artificial agents when they help humans. All positive and negative moral actions have normative components that are culturally specific. There are many norms and expectations that may differ between actor types. For instance, artificial agents are expected to make decisions that sacrifice one for the good of many, while humans are expected not to (Malle et al., 2015).

Our short format scenarios leave the physical embodiment of the artificial agent up to the imagination of the participant. For instance, is the agent a large intimidating robot, a small desktop device, or a smartphone? Is it anthropomorphic or machine-like? Each of these has a very different social presence and will be perceived very differently by the people around it. Differences in norms between these physically different kinds of artificial agents may alter mind perception and trust (Waytz, Heafner, & Epley, 2014; Zhao, Phillips, & Malle, under review), but further research could include moral attribution as well.

Finally, research to determine the degree to which the liminality of agency is a factor in blame and attribution would build greatly upon the work of this research. If the perception of this liminality could be changed through either physical or behavioral changes in the agents that would greatly strengthen claims about the relationship of agency to blame attribution in artificial agents.

Conclusions

These results have wide ranging implications for the implementation of artificial agents and the ways users might think of them. These agents are only liminally minded and partially agentic, which gives them some shielding from blame, reflecting it instead to those responsible for creating, placing, or maintaining the agent. However, this does not exempt the agent from blame when direct harm is perpetrated. In that situation blame is attributed similarly to a human performing the same action, likely due to the agent-neutral reasons that harm is considered wrong in the first place. Therefore, an artificial agent which perpetrates a harm violation may take some blame itself instead of passing all blame to its manufacturer.

Artificial agents may be perceived differently in moral actions dealing with group hierarchies as evidenced by reduced attribution of subversion and oppression violations to them. Practically this may mean that agents can subvert authority or oppress others without a moral backlash by society. If machines, which are not typically part of human groups and hierarchies, are included in them they may benefit from moral freedom that humans in equivalent positions do not have. This may present avenues for these violations to be perpetrated with lessened repercussions.

These ideas raise the question about injunctive norms as well: *should* this be the way artificial agents are judged? An artificial agent which acts subversively may receive less blame, preventing that blame from being passed on to a party which deserves it. This may be supported by the agent's functional limitations or on the public's perception of robots as highly logical and utilitarian. Conversely an artificial agent which is assigned blame for direct harm may shield its manufacturer from scrutiny serving as a scapegoat. Blame for an agent could be associated with a glitch in a single unit, poor maintenance, or even actions taken by the victim.

We hope this research will contribute to a new direction in moral foundations theory that considers types of morality across types of moral agents. While we did not discuss in detail implications outside of the social psychological results, understanding how humans may similarly and differently perceive the immorality of robots, artificial intelligence, and machines has implications for computer scientists, technology designers, ethicists, and policy makers. In the future, artificial agents will engage in a broader range of ethically-charged actions, and it is paramount to have a better understanding of human perception of them.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-19-1-0246 to Dr. Daniel B. Shank. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Special thanks for feedback from Courtney Stefanik, Carson Arnold, and Abigail Wilson.

Appendix A. Scenarios in Study 2's human condition compared to Clifford et al. (2015)

We conducted 16 z-tests between the proportions of respondents that selected the expected moral foundation in the human conditions compared to analogous ones in Clifford et al. (2015). Thirteen z-tests indicated no statistical difference between our Study 2 human condition and Clifford et al.'s scenarios in terms of the expected moral foundation choices. The three that were significantly different showed our study 2 respondents perceiving less liberty violation in the betrayal scenario *Dictator* (66.6% in Clifford et al.; 43.9% in study 2, $z = 2.055$, $p = .039$), less purity violation in the sanctity scenario *Sex* (73.3% in Clifford et al.; 46.8% in Study 2, $z = 2.294$, $p = .022$) and less liberty in the oppression scenario *Colors* (62.5% in Clifford et al.; 29.8% in Study 2, $z = 2.752$, $p = .006$).

All three of the scenarios had changes to their main actors in our Study 2 compared to the original (*Dictator*: a Hollywood star to a university professor; *Sex*: a homosexual to "someone"; *Color:* a pastor to an HR coordinator; see Table 1) which may have changed perception of the moral foundation. However, on closer examination, all three of these had extremely high rates of people perceiving them as "not wrong" (*Dictator*: 38.8%; *Sex*: 27.8%; *Colors*: 57.4%) suggesting the change in actors mainly changed the wrongness of them, not the foundation they violated. When removing the not wrong option and recalculating the percentages, all three of those scenarios exceed the 60% threshold used by Clifford et al. (2015) for inclusion.

References

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., et al. (2019). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour, 4*, 134–143.

Banks, J. (2020). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, 1–18.

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

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

Chokshi, N. (2018). Amazon knows why alexa was laughing at its customers. In *The New York times*. https://www.nytimes.com/2018/03/08/business/alexa-laugh-amazon-echo.html.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods, 47*(4), 1178–1198.

D'Errico, F., & Paciello, M. (2018). *Online moral Disengagement and Hostile Emotions in Discussions on Hosting Immigrants*. Internet Research.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor: St. Martin's Press*.

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & Society, 35*, 795–809.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., et al. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Elsevier.

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

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130.

Longoni, C., & Cian, L. (2020). Artificial intelligence in utilitarian vs. Hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*. https://doi.org/10.1177/0022242920957347

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). Ai in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In I. A. Ferreira,

J. S. Sequeira, G. S. Virk, E. E. Kadar, & O. Tokhi (Eds.), *Robots and well-being* (pp. 111–133). Springer.

Malle, B. F., & Scheutz, M. (2016). Inevitable psychological mechanisms triggered by robot appearance: Morality included?. In *2016 AAAI spring symposium series*.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117–124). ACM.

McManus, R. M., & Rutchick, A. M. (2018). *Autonomous Vehicles and the Attribution of moral Responsibility*. Social Psychological and Personality Science, 1948550618755875.

Miller, G. (2012). Drone wars. *Science, 336*(6083), 842–843.

Neff, G., & Nagy, P. (2016). Automation, algorithms, and Politics| talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication, 10*, 17.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Rao, S., Griffis, S. E., & Goldsby, T. J. (2011). Failure to deliver? Linking online order fulfillment glitches with future purchase behavior. *Journal of Operations Management, 29*(7–8), 692–703.

Ridge, M. (2005). Reasons for action: Agent-neutral vs. Agent-relative. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Stanford University.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The cad triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 76*(4), 574–586.

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin, 41* (8), 1147–1163.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411.

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

Shank, D. B., & Gott, A. (2020). Exposed by ais! People personally witness artificial intelligence exposing personal information and exposing people to undesirable content. *International Journal of Human-Computer Interaction, 36*(17), 1636–1645.

Shank, D. B., North, M., Arnold, C., & Gamez, P. (2021). Can mind perception explain virtuous character judgments of artificial intelligence? *Technology, Mind, and Behavior, 2*(3).

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! An interaction with a cheating robot. In *2010 5th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 219–226). IEEE.

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLoS One, 12*(7), Article e0180952.

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

Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. WW Norton & Company.

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

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*, 103870.

Zhao, X., Phillips, E., & Malle, B.F. (under review). How People Infer a Humanlike mind from a Robot Body.