Can Mind Perception Explain Virtuous Character Judgments of Artificial Intelligence?

Daniel B. Shank¹, Mallory North¹, Carson Arnold², and Patrick Gamez³

¹ Department of Psychological Science, Missouri University of Science and Technology

² Department of Chemistry, University of Missouri

³ Department of Arts, Languages & Philosophy, Missouri University of Science and Technology

People tend to attribute less of a virtuous or unvirtuous characteristic to artificial intelligence (AI) agents compared to humans after observing a behavior exemplifying that particular virtue or vice. We argue that this difference can be explained by perceptions of experiential and agentic mind. Experiential mind focuses on one's emotions, sensations, and past experiences, whereas agentic mind focuses on one's intentions, capacity for action, and behaviors. Building on person-centered morality, virtue ethics, and mind perception research, we argue that both agentic and experiential mind are possible mediators of behavior-to-character attributions. We conducted two experiments ($n = 613$, $n = 584$) using vignette scenarios in the virtue ethics domains of truth, justice, fear, wealth, and honor where we manipulated the actor to be an AI or human and the behavior to be virtuous or unvirtuous. As expected, we found that the character judgments of virtues and vices are weaker for AIs compared to humans. This character judgment difference is mediated by both experiential and agentic mind with a larger mediation effect for experiential mind compared to agentic mind. Exploratory analyses revealed differences in character and experiential mind based on the virtue domain.

Keywords: mind perception, person-centered morality, character judgments, machines, artificial agents

Supplemental materials: <https://doi.org/10.1037/tmb0000047.supp>

Artificial intelligence (AI) and intelligent machines used to autonomously drive a vehicle, interpret medical scans, or perform military missions may behave in ways that have ethical implications (Cummings, 2017; Safadi et al., 2015; Topol, 2019). Unethical behavior enacted by AIs is typically judged as immoral (Malle et al., 2019; Shank, DeSanti, et al., 2019), yet little is known about how perceptions of immoral behavior relate to judgments that an AI possesses virtuous characteristics such as being stingy versus generous or honest versus dishonest. Virtue ethics and person-centered morality (PCM) both argue that attributions of these virtuous characteristics can be critical for impression formation, understanding other's intentions, and moral personhood. Therefore, understanding how virtuous character judgments differ for AIs

compared to humans extends the research on morality of AIs beyond behavioral judgments.

In previous research people attributed virtuous character to both humans and artificial agents across types of virtue domains; however, humans receive stronger character judgments compared to AIs—both for virtuous and unvirtuous judgments (Gamez et al., 2020). This study considers mind perception as a potential mechanism to explain this difference, leading to the central research question: How does mind perception change character judgments of humans versus AIs? Perceiving an AI or human's level of mind—specifically in terms of an agentic, intentional mind, and an experiential, feeling mind—has been linked previously to moral behavior judgments (Bigman et al., 2019; Gray & Wegner, 2009, 2012), but

Action Editor: Danielle S. McNamara was the action editor for this article.

ORCID iDs: Daniel B. Shank <https://orcid.org/0000-0002-3746-2407>; Patrick Gamez <https://orcid.org/0000-0002-9353-1817>.

Funding: This research was partially supported by the Army Research Office under Grant Number W911NF-19-1-0246. This work was also funded by a grant from Missouri University of Science and Technology's Center for Science, Technology, and Society to Daniel B. Shank and Patrick Gamez.

Acknowledgments: 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. We would also like to thank Hanne Watkins, Courtney Stefanik, Mariter White, Timothy Maninger, Emily Swaters, and Abigail Wilson for their comments on this article.

Conflicts of Interest: There are no perceived or potential conflicts of interest.

Data Availability: The data and study materials for this study are available at <https://osf.io/dwv2q/>.

Open Access License: This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC-BY-NC-ND). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Disclaimer: Interactive content is included in the online version of this article.

Contact Information: Correspondence concerning this article should be addressed to Daniel B. Shank, Department of Psychological Science, Missouri University of Science and Technology, 500 W. 14th Street, Rolla, MO 65409, United States. Email: shankd@mst.edu

not to virtuous character judgments. A secondary research question explores differences across type of virtues and vices: How do character judgments and mind perception for humans and AIs differ across virtue domains and for virtues and vices?

The paper is organized by first discussing moral character of AIs, then introducing person-centered morality and the virtue domains, before discussing mind perception and presenting hypotheses. Two experiments then investigate the research questions and test the hypotheses, with a final discussion considering the limitations and implications of the findings.

Moral Character and Behavior of Artificial Intelligence

Judgments of the character of other humans are crucial to the way people navigate through the world. That is, if people are better at figuring out who is just, generous, and honorable, they can better fulfil their own goals, enact better relationships, find moral role models, and predict others' behavior. Specifically, moral character shapes the development of perception of others (Goodwin, 2015; Goodwin et al., 2014) and influences social desirability (Landy & Goodwin, 2015). Specifically, facets of moral character can predict one's cooperativeness (Hilbig et al., 2014; Zettler et al., 2013), generosity (Cohen et al., 2014; Hilbig & Zettler, 2009), and prosociality of future behaviors (O'Boyle et al., 2012; Moore et al., 2012).

Moral character may also be important for evaluations of AI agents as these are similar to evaluations of humans (Banks et al., 2021). Increasingly, people label AIs and machines in more human-like terms including character judgments when they are highly anthropomorphic (Kiesler et al., 2008; Li & Sung, 2021; Schroeder & Epley, 2016; Waytz et al., 2014) or engage in unethical behavior (Bigman et al., 2020; Shank & Gott, 2020). If the process of making these character judgments differs from humans, the attribution of virtuous characteristics could have implications for explaining other human-agent differences across domains. For example, if people are less likely to judge AIs as just, then that attribution might explain why people have less trust in autonomous vehicles (Waytz et al., 2014) and medical AIs (Bigman & Gray, 2018; Yokoi et al., 2020).

Much of the research on AI morality focused not on judgments of character, but of behavior. These judgments occur in a variety of domains including medical decisions (Bigman & Gray, 2018), real-world discrimination (Shank & DeSanti, 2018), and life or death decisions based on moral dilemmas like the trolley problem (Malle et al., 2015; Voiklis et al., 2016). Usually judgments of the severity of the immoral behavior are greater for humans compared to AIs (Malle et al., 2019; Shank, DeSanti, et al., 2019) with mind perception as one potential mediator of this process (Gray & Wegner, 2012; Shank & DeSanti, 2018). To consider judgments of character more fully, we turn to person-centered morality.

Person-Centered Morality

The theory of person-centered morality (PCM) asserts that moral cognition aims predominantly at character evaluations as opposed to the judgment of specific behaviors or outcomes (Landy & Uhlmann, 2018; Pizarro & Tannenbaum, 2012; Uhlmann et al., 2015). In social interactions, it is important to identify others' intentions toward us (Abele & Wojciszke, 2007; Pizarro & Tannenbaum,

2012; Wojciszke et al., 1998), and moral character is the primary indicator of these intentions as it is largely responsible for impression formation (Landy & Goodwin, 2015). PCM is rooted in the philosophical theory of virtue ethics, which is an Aristotelian moral theory in Western philosophy (Kraut, 2001). In brief, virtue ethicists argue that the primary object of moral importance is character—that is, a set of reliable dispositions—and these are reflected in situationally enacted behavior.

PCM also accepts this premise and seeks to determine that behaviors are most revealing of one's moral character including when behavior and character judgments differ. Good behaviors are typically less indicative of moral character than their negative counterparts because those with poor moral character are capable of acting positively as a means of deception (Landy & Uhlmann, 2018; Skowronski & Carlston, 1989). This can lead to character judgment that is worse than behavioral judgment for a single act (Pizarro & Tannenbaum, 2012). An example of this *act-person dissociation* occurs when people judge a racist manager who slurs at a coworker as having a worse moral character than one who punches a coworker, even though the act of punching is judged as more immoral than the slur (Uhlmann et al., 2014). Thus, there is mounting evidence from PCM that the virtuosity of actions is not necessarily aligned with the moral character of the individual (Tannenbaum et al., 2011; Uhlmann et al., 2009). These show the utility of understanding character judgments apart from judgments of moral behavior in humans and suggest a similar distinction may be useful for AIs.

Virtue Domains

While PCM research tends to consider only a simple positive and negative judgment of overall character, virtue ethics considers character judgment as constituted by multiple discrete virtues and vices (McKinnon, 1999, pp. 27–51). Five of the most prominent domains include *truth, justice, fear, wealth, and honor* with their corresponding virtues (honesty, justice, courage, generosity, and humility) and vices (dishonesty, unfairness, cowardice, stinginess, and pride). Prior research has considered what factors lead to one possessing specific virtues including honesty (Whetstone, 2003), fairness (Bragues, 2005; Hackett & Wang, 2012; Riggio et al., 2010), courageousness, generousness, and humility (Barker et al., 2003; Chun, 2005) with less research focusing on factors leading to unvirtuous characteristics.

It is possible that people are reluctant to make certain character judgments for AIs that they would more readily make toward humans (e.g., generous or stingy), whereas character judgments in other virtue domains may not differ as much (e.g., just or unjust). For example, people might expect Amazon's Alexa—as a personal assistant that retrieves and organizes information—to be honest and just but might not expect Alexa to exhibit the virtues of generosity or courage. Considering multiple virtue domains allow an exploration of how specific character judgments may differ for AIs compared to humans. Since multiple virtue domains have not been used in PCM research on humans, this may also extend PCM in a more fine-grained way may enable better recognition of their intentions, and thus their probable future actions (Landy & Goodwin, 2015). We have no a priori evidence or theory to expect a difference in judgments of AI character by virtue domain. As such we examine them in an exploratory analysis.

Mind Perception and Morality

Our central research question regards the extent to which mind perception explains the relationship of character judgments to AI and human identity. Gray et al. (2007) have proposed observers perceive two primary dimensions of mind: an agentic and an experiential component. Agentic mind is the ability to act, evaluate, learn, and make decisions, while experiential mind is the ability to have feelings, sensations, and remember past experiences. When one attributes agentic mind to another, the actions of the perceived are viewed as purposeful and potentially moral. Likewise, if people attribute experiential mind to another, they perceive them as having an inner life and feeling pain, making actions toward them potentially moral (Waytz et al., 2010). Therefore if an action is harmful, perceiving agentic mind to the perpetrator and experiential mind to the victim leads to the perception of the morality of the act (Gray & Wegner, 2009; Schein & Gray, 2018).

In general, people attribute a moderate level of agentic mind and little, if any, experiential mind to AIs, robots, and machines (Gray et al., 2007; Gray & Wegner, 2012). However, anthropomorphic features, such as looking like a human, can increase the levels of perceived mind (Martini et al., 2016). This increase, predominantly in perceptions of experiential mind (Appel et al., 2020; Gray & Wegner, 2012), can lead observers to feel uneasy when a robot's appearance is in *the uncanny valley*—that is near, but noticeably different from a human (Gray & Wegner, 2012). Speed of movement near that of humans (Morewedge et al., 2007) and humanlike voices (Schroeder & Epley, 2016) can also increase the perceived mind of AIs.

Mind perception has a clearly discernible effect on how humans judge moral aspects of AI. Agentic mind increases assessments of the permissibility of an AI making a moral decision (Bigman & Gray, 2018). Greater mind perception of AIs relates to higher levels of attributed wrongness when making a moral decision (Malle et al., 2019) and more blame when committing a moral wrong (Shank & DeSanti, 2018). In instances of human-AI collaboration, perceptions of greater agentic and experiential mind can lead to higher levels of cooperation (De Melo et al., 2014), but people also ascribe mind in self-serving ways based on cooperation, competition, and outcomes (Lefkeli et al., 2020). Perpetrating a moral wrong necessitates perceiving the perpetrator as having at least some level of agentic mind, whereas being the victim of a moral wrong necessitates perceiving the victim as having some level of experiential mind (Gray et al., 2012). This link goes both ways with mind perceptions changing based on the immorality of a behavior (Gray & Wegner, 2009; Ward et al., 2013) and having at least partially minded perpetrators and victims as a prerequisite to classifying an act as immoral (Gray et al., 2014; Schein & Gray, 2018).

From Mind Perception to Character Judgments

We argue that both dimensions of mind perception may mediate differences in the attribution of virtuous characteristics to AIs. In a broad sense, perceiving an AI as having a more agentic or experiential mind brings it closer to the perceived level of a human mind, thereby enabling the type of judgments made of humans.

Specifically, agentic mind is tied to action, the intentionality, and purpose behind that action (Gray et al., 2007) and moral behavior

(Gray et al., 2012). Some scholars argue that without an actor with intention, a behavior cannot be moral or immoral, and conversely a moral or immoral behavior implies an intentional actor (Gray & Wegner, 2009; Schein & Gray, 2018). Applied to PCM, those who see moral behavior indicative of a virtue or vice make a judgment of character to the actor. However, if this same behavior is performed by an actor perceived to have little or no agency, it may be dismissed as neither intentional nor reflecting the actor's character. Furthermore, AIs are generally perceived as having less agentic mind than humans (Gray et al., 2007; Wegner & Gray, 2017). Therefore, this suggests,

Hypothesis 1: Perceived agentic mind is positively related to greater (un)virtuous character judgments.

Hypothesis 2: Differences between humans and AIs in (un) virtuous character judgments are mediated by agentic mind perception.

Experiential mind relates to one's capacity for internal states such as emotions, pain, and experiences, making it more difficult for observers to perceive it directly (Wegner & Gray, 2017). Character is a similar construct primarily focused on one's internal nature that can manifest externally (e.g., through behavior). Therefore, experiential mind and virtuous character may overlap in respect to how they are perceived. Yet, they are quite different concepts. Experiential mind's capacity to experience sensations does not require those sensations to have occurred, although their occurrence, especially in the case of causing pain or harm, can increase one's perception of experiential mind (Gray & Wegner, 2009; Ward et al., 2013). In contrast, PCM and virtue ethics posit that observation of behavior allows observers to infer a specific virtuous or unvirtuous character trait, not simply the capacity for such a trait.

However, perceiving a greater capacity for mental life, and related experiences, should be a prerequisite to being able to attribute virtuous character. For a truthful statement to indicate honesty, the listener must assume that the speaker has the internal mental capacity to perceive the truth versus a lie and chooses the truthful option, despite potential negative repercussions. This means that character judgments imply some level of subjective, affective contemplation, beyond a rational weighing of alternatives or following rules, and suggest a link with experiential mind. Within the dimensions of mind perception, humans are the prototypical beings associated with high experiential mind, whereas AIs are expected to have low experiential mind (Gray & Wegner, 2012; Shank, Graves, et al., 2019; Wegner & Gray, 2017). Thus, we predict,

Hypothesis 3: Perceived experiential mind is positively related to (un)virtuous character judgments.

Hypothesis 4: Differences between humans and AIs in (un) virtuous character judgments are mediated by experiential mind perceptions.

Study 1

Study 1 is an experiment on participants' character judgments and mind perceptions of humans or AIs behaving virtuously or unvirtuously in a scenario. The study was a $2 \times 2 \times 5$ factorial design with identity (AI or human) as a between subjects factor, virtue domains

(*truth, justice, fear, wealth, or honor*) as a within subjects factor and behavior (virtuous or unvirtuous) as a repeated within subjects factor. Specifically, participants were randomly assigned to either receive all AI or all human scenarios, and then randomly cycled through five scenarios, one covering each virtue domain. A randomly selected behavior (virtuous or unvirtuous) was included as part of each scenario.

Stimuli

We used 15 scenarios, three from each virtue domain, which described JD who is either a human or AI engage in a virtuous or unvirtuous behavior. Previous research using these scenarios confirmed that each produced judgments of the expected virtuous or unvirtuous characteristic for humans but also showed some, albeit weaker, judgments of these characteristics for AIs (Gamez et al., 2020). This previous study did not compare across virtue domains, nor did it include any measures of mind; therefore, neither the hypotheses nor the exploratory analyses have been previously tested.

The scenarios average 74.7 words and involve a range of situations including JD giving business presentations, creating a group memo, serving as a financial account analyst, commanding troops, or approving inmate requests (see the Appendix for all scenarios). An example from the *wealth* domain is

JD is [an AI program that/a full-time musician who] makes ample money by creating unique albums for a record label. During a large hurricane, the local music school was partially destroyed and has a fundraising event to raise money to rebuild. After learning of this news, JD [*generous*: helps with the fundraiser and donates funds to help rebuild, instead of spending money on new equipment/*stingy*: ignores the fundraiser and spends money on new musical equipment].

Measures

Immediately after the scenario, the participants were asked to make a character judgment (i.e., “How would you characterize JD?”) along a 7-point scale from “completely” the vice (*dishonest, unjust, cowardly, stingy, or conceited*) to “completely” the virtue (*honest, just, courageous, generous, or humble*). The points were tagged as “somewhat” (coded ± 1), “quite” (± 2), and “completely” (± 3) with a neutral option (e.g., “Neither honest nor dishonest”) in the middle (0) with the positive number coding being the expected virtue or vice. An eighth “Not Applicable/None of these Apply” option was always presented as well.

Based on research that had applied mind perceptions to AIs, robots, computer agents, and machines (De Melo et al., 2014; Gray & Wegner, 2012; Morewedge et al., 2007; Shank & DeSanti, 2018; Stafford et al., 2014), we developed 16 items that could reasonably apply to AIs (Table 1). These were presented in a random order with a 5-point scale from “*Completely Disagree*,” to “*Completely Agree*.” A principal components analysis with a varimax rotation with a Kaiser normalization revealed two components with 10 items uniquely loading high ($>.5$; Table 1: bolded) on Component 1 (experiential mind), five items uniquely loading high on Component 2 (agentic mind), and one item loading high on both (Table 1). Mind scores were regressed based on these loadings, and therefore have a mean of 0 and standard deviation of 1.

A few unused additional questions about morality were asked after the primary measures, and demographics were assessed at the end of the survey, with details available at Shank et al. (2021).

Participants and Exclusions

Our 614 U.S. Amazon Mechanical Turk participants (67% white, 59% male, mean age of 36) completed the survey and one was excluded for finishing in under 2 min. Each participant was presented five scenarios, for a potential total of 3,065 character rating cases. In 3.8% (117) of the cases participants selected the not applicable option. Like previous research with these scenarios (Gamez et al., 2020), more *not applicable* selections were made for AIs (97) compared to humans (20). We excluded these as they are cases of respondents choosing to not make character judgments, therefore are not appropriate for testing our hypotheses. Additionally, there were four cases the character judgments were left blank which were also excluded. In 17.4% (534) of the cases, respondents rated the vice using a virtue rating or the virtue using a vice rating and therefore are excluded.¹ After all exclusions, 2,410 cases remain that are most appropriate to test our hypotheses. The research was approved by Missouri University of Science and Technology’s institutional review board. The data and study materials for this study are available at Shank et al. (2021).

Study 1 Results

Analyses on Character Judgments (Exploratory)

We first examine which experimental conditions influence character judgments. To do so, we conduct a full-factor generalized linear mixed model—structured based on our data with an identity link function and a diagonal covariance type—to predict character judgments (Table 3: Model 1). As expected, character judgments of AIs ($M = 1.72$, $SD = 1.19$; Table 2) were weaker than those of humans, $M = 1.91$, $SD = 1.04$; $F(1, 2390) = 18.115$, $p < .001$.² Therefore, we replicated the effects found in our previous research (Gamez et al., 2020).

Additionally, people made stronger character judgments for virtuous behavior ($M = 1.89$, $SD = 1.14$) than unvirtuous behavior, $M = 1.73$, $SD = 1.10$; $F(1, 2390) = 13.225$, $p < .001$. The virtue domain was also significant, $F(4, 2390) = 14.387$, $p < .001$, with a post hoc LSD test showing character judgments for *truth* ($M = 1.96$,

¹ Over three-fourths of those cases involved the respondent mixing the virtue and vice at least twice, suggesting some participants were either not actively paying attention or rushing through the survey. Others may have genuinely misinterpreted the behavior in the scenario, or believed the virtuous behavior was unvirtuous or vice versa. Since this is a large number of cases and the exact reason participants rated the virtue opposite of expected could not be determined, we also conducted the analysis with these included. When these cases are included and coded as part of the virtue-to-vice scale as negatives, it weakens the possibility of finding hypothesized results by adding noisy data. Nevertheless, *Hypotheses 1, 3, and 4* are supported with this data included. All hypothesized results and analyses with results that change significant levels are noted in footnotes and refer to these as mixed cases.

² When the mixed cases are included, the effect of identity is eliminated, $F(1, 2924) = .602$, *ns*, and replaced with a three-way identity $\times$ virtuous behavior $\times$ virtue domain effect, $F(4, 2924) = 2.560$, $p = .037$. The lack of a first-order identity effects and the presence of the three-way interaction effect remain in Models 4 and 5.

Table 1*Two Component Varimax Rotated Solution for Mind Perception Items for Studies 1 and 2 (Loadings > .5 Are in Bold)*

Items ^a	Study 1		Study 2	
	Component 1: Experiential mind	Component 2: Agentic mind	Component 1: Experiential mind	Component 2: Agentic mind
JD can experience emotional pain or pleasure	0.889	0.198	0.899	0.237
JD can feel distress	0.859	0.207	0.873	0.245
JD has a personality	0.852	0.267	0.851	0.321
JD can feel anticipation	0.852	0.260	0.853	0.293
JD can recognize sensations	0.852	0.239	0.876	0.285
JD has desires	0.841	0.281	0.828	0.342
JD has beliefs	0.835	0.308	0.807	0.373
JD can recognize emotions	0.826	0.288	0.863	0.294
JD can have experiences	0.768	0.390	0.731	0.466
JD has a mind of its own	0.727	0.361	0.709	0.426
JD has intentions	0.581	0.526	0.556	0.586
JD can remember the past	0.493	0.545	0.458	0.563
JD can reason	0.462	0.601	0.510	0.609
JD seeks continued functioning	0.402	0.576	0.331	0.642
JD can plan actions	0.181	0.825	0.236	0.820
JD can act in order to meet its goals	0.061	0.843	0.103	0.841

^a Ordered by component 1 loadings from Study 1.

$SD = 1.09$) and *justice* ($M = 2.02$, $SD = 1.09$) were greater than the three other domains (*truth* vs. *fear*: $p < .001$; *truth* vs. *wealth*: $p < .001$; *truth* vs. *honor*: $p = .034$; *justice* vs. *fear*: $p < .001$; *justice* vs. *wealth*: $p < .001$; *justice* vs. *honor*: $p = .002$). It also showed that character judgments for *wealth* ($M = 1.72$, $SD = 1.13$) and *honor* ($M = 1.81$, $SD = 1.10$) were each greater than the *fear* domain ($M = 1.54$, $SD = 1.15$; *wealth* vs. *fear*: $p = .012$; *honor* vs. *fear*: $p < .001$).

Additionally, there was an interaction effect between behavior type and virtue domain on judgments of character, $F(4, 2390) = 3.333$, $p = .010$, where virtuous versus unvirtuous character judgments differed by domain, with the *fear* domain differing the most from the others: *truth* (virtuous minus unvirtuous judgments: $-.02$), *justice* (.15), *fear* (.47), *wealth* (.03), and *honor* (.15; Figure S1). These analyses show that while there was some variation by virtue domain and by behavior, these did not significantly interact with identity in Study 1. That is, this analysis indicates no evidence of a

different process for AI versus human character judgments based on virtue domains, nor by positive or negative behavior.

Analyses of Mind Perception (Exploratory)

The experimental condition's influence on mind perception is examined by conducting analogous full-factor generalized linear mixed models predicting agentic (Table 3: Model 2) and experiential mind (Model 3).

As expected, AIs ($M = -0.04$, $SD = 1.15$; Table 2) were perceived to have less agentic mind than humans, $M = 0.04$, $SD = 0.81$; $F(1, 2390) = 5.978$, $p < .001$, Table 3: Model 2.³ Additionally, virtuous behavior ($M = 0.13$, $SD = 0.95$) led to perceiving more agentic mind than unvirtuous behavior, $M = -0.13$, $SD = 1.04$; $F(1, 2390) = 41.437$, $p < .001$. There were no other effects on agentic mind.

All three experimental factors had significant effects on experiential mind. As expected, AIs ($M = -0.65$, $SD = 0.96$) were perceived to have less experiential mind than humans, $M = 0.66$, $SD = 0.47$; $F(1, 2,390) = 1845.955$, $p < .001$. Virtuous behavior ($M = 0.03$, $SD = 1.02$) led to perceiving more experiential mind than unvirtuous behavior, $M = -0.03$, $SD = 0.98$; $F(1, 2390) = 13.728$, $p < .001$. Virtue domain also led to significantly different levels of experiential mind perception, $F(4, 2390) = 9.796$, $p < .001$. An LSD post hoc test indicated that domains of *fear* ($M = 0.04$, $SD = 1.01$), *wealth* ($M = 0.04$, $SD = 1.00$), and *honor* ($M = 0.12$, $SD = 0.94$) had significantly higher experiential mind than *truth* ($M = -0.07$, $SD = 1.02$; *truth* vs. *fear*: $p = .022$; *truth* vs. *wealth*: $p = .023$; *truth* vs. *honor*: $p < .001$) and *justice* ($M = -0.13$, $SD = 1.01$; *justice* vs. *fear*: $p < .001$; *justice* vs. *wealth*: $p < .001$; *justice* vs. *honor*: $p < .001$). No two-way interactions were significant; however, a three-way interaction effect was significant,

³ When the mixed cases are included, the identity main effect drops to marginal significance, $F(1, 2924) = 3.528$, $p = .06$, and previously marginally significant identity $\times$ domain effect becomes significant, $F(4, 2924) = 3.015$, $p = .017$.

Table 2*Variable Means (Standard Deviations) Across Experimental Conditions for Study 1*

Condition	N	Character attribution	Agentic mind	Experiential mind
All	2,410	1.81 (1.12)	.00 (1.00)	.00 (1.00)
Identity				
Human	1,199	1.91 (1.04)	.04 (.81)	.66 (.47)
AI	1,211	1.72 (1.19)	-.04 (1.15)	-.65 (.96)
Behavior				
Virtuous	1,228	1.89 (1.14)	.13 (.95)	.03 (1.02)
Unvirtuous	1,182	1.73 (1.10)	-.13 (1.04)	-.03 (.98)
Domain				
Truth	485	1.96 (1.09)	-.01 (1.02)	-.07 (1.02)
Justice	492	2.02 (1.09)	-.03 (.99)	-.13 (1.01)
Fear	471	1.54 (1.15)	.03 (.99)	.04 (1.01)
Wealth	481	1.72 (1.13)	-.03 (.99)	.04 (1.00)
Honor	481	1.81 (1.10)	.05 (.97)	.12 (.94)

Note. AI = artificial intelligence.

Table 3*Generalized Linear Mixed Models Predicting Character Attributions and Mind Perception for Study 1 (Fs Reported)*

Variable	df	Model 1	Model 2	Model 3	Model 4	Model 5
		Character	Agentic mind	Experiential mind	Character	Character
Identity (AI vs. human)	1	18.115***	5.978*	1845.955***	15.763***	1.239
Behavior (virtuous vs. unvirtuous)	1	13.225***	41.437***	13.728***	5.742*	10.006**
Domain	1	14.387***	1.059	9.796***	15.459***	17.040***
Identity × Behavior	1	2.616	.097	.601	2.514	2.335
Identity × Domain	4	.718	2.342†	1.162	1.276	.892
Behavior × Domain	4	3.333**	.481	.907	3.706**	3.684**
Identity × Behavior × Domain	4	1.332	.634	2.402*	1.653	1.863
Agentic mind (covariate)	1				96.272***	
Experiential mind (covariate)	1					43.277***

Note. $N = 2,410$. AI = artificial intelligence.

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

$F(4, 2390) = 2.402, p = .048$.⁴ No single cells showed an extreme difference (Figure S2), making this effect difficult to meaningfully interpret. One potential explanation is that unvirtuous conditions typically produced lower experiential mind than virtuous conditions, with the exceptions of AIs in the *justice* domain and humans in the *fear* domain.

Next, the mean value of experiential mind (.66) was significantly higher than agentic mind (.04) for humans, $t(1199) = 27.15, p < .001$, while for AIs the mean experiential mind (−.65) was significantly lower than agentic mind, −.04; $t(1210) = 13.02, p < .001$. This indicates that AIs and humans differ more from each other in experiential mind (1.31) compared to agentic mind (0.08). In sum, these analysis show that while there was some variation by virtue domain and by behavior on mind perception, for these did not significantly interact with AI identity, save the one three-way interaction on experiential mind.

Hypotheses: Mind Perception as a Mediator of AI Versus Human Character Judgments

The hypotheses are tested with two analogous full-factor generalized linear mixed models predicting character judgment similar to Model 1, but with agentic mind (Table 3: Model 4) and experiential mind (Model 5) as covariates. Respectively, each is weakly but positively associated with character judgment (agentic mind: $r = .194, p < .001$; Experiential Mind: $r = .144, p < .001$) and also significant in the models (agentic mind: $F(1, 2389) = 96.272, p < .001$; experiential mind: $F(1, 2389) = 43.277, p < .001$,⁵ which control for the experimental factors. Therefore, these fully support *Hypotheses 1 and 3*, that character judgments are related to mind perceptions.

We test for mediation by conducting a mediation analysis using Hayes Process Macro version 3 (Hayes, 2017) with 10,000 bootstrap samples to test whether mind perception mediated AI versus human identity's effect on character judgments. Both experiential and agentic mind were entered as potential mediators. AI identity, compared to human identity, negatively influenced both agentic mind (−.088, $SE = .041, p = .031$; Figure 1)⁶ and experiential mind (−1.309, $SE = .031, p < .001$), and both of these had a positive effect on character judgments (agentic: .220, $SE = .022, p < .001$; experiential: .183, $SE = .029, p < .001$). There were significant indirect effects of AI on character judgments through both agentic

mind (−.019; CI: −.038, −.002) and experiential mind (−.240; CI: −.320, −.162)⁷, thereby supporting *Hypotheses 2 and 4*. Notably, while still a technically significant mediation, agentic mind's substantive contribution to predicting character judgments was only 7.9% of experiential mind's contribution.

Study 2

Study 2 replicates Study 1 and is identical and saved several methodological improvements. First, the only options for the character judgment question (i.e., “How would you characterize JD?”) were in relation to the presented virtue or vice (i.e., not both as in Study 1), again at three levels (“somewhat” coded 1, “quite” 2, and “completely” 3), with a neutral option (coded 0; e.g., “Neither honest nor dishonest”) and a “Not Applicable/None of these Apply” option (excluded). Second, an attention check question, (i.e., “If you are paying attention do not select any answers to this question.”) was included once after the mind items. Third, Study 2 was conducted using respondents from Prolific.co, a high-quality participant pool website designed for online research. Fourth, we preregistered Study 2 at <https://osf.io/dwv2q/>, where the data and study materials are also available.

Participants and Exclusions

Of the 604 U.S. Prolific participants (44% male, 54% female, 72.2% White, 33.4 years old) who completed the survey, none were excluded for finishing in under 2 min. However, 20 who failed the attention check and were excluded, leaving 584 participants.

Of the 2,920 character ratings (e.g., 584 × 5 virtue domains) in 11.0% (320) of the cases participants selected the *not applicable* option which was included so that no one was forced to make a

⁴ It is only marginally significant when mixed cases are included, $F(4, 2924) = 21.95, p = .067$.

⁵ With mixed cases included both still significant support the hypotheses, agentic mind in Model 4: $F(4, 2923) = 81.95, p < .001$; experiential mind in Model 5: $F(4, 2923) = 9.44, p = .002$.

⁶ With mixed cases included both still significant support the hypotheses, Agentic mind in Model 4: $F(4, 2923) = 81.95, p < .001$; experiential mind in Model 5: $F(4, 2923) = 9.44, p = .002$.

⁷ With mixed cases included the indirect effect of agentic mind was the same but was no longer significant (−.019; CI: −.040, .002), failing to support *Hypothesis 2*, while the effect of experiential mind was still significant (.169; CI: .060, .261), supporting *Hypothesis 3*.

Figure 1

Mediation Model of Identity (I = Artificial Intelligence; 0 = Human) on Character Attributions for Study 1

Experiential Mind Indirect Effect: -.240 [CI: -.320, -.162]

Agentic Mind Indirect Effect: -.019 [CI: -.038, -.002]

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

character judgment of an AI. Like Study 1, more *not applicable* selections were made more for AIs (258) compared to humans (62). We exclude these cases and one additional case where character rating was left blank, leaving 2,599 cases for the analysis.

Mind Measures

A principal components analysis with a varimax rotation with a Kaiser normalization produced a solution with 10 items uniquely loading high ($>.5$; Table 1: bolded) on Component 1 (experiential mind), four items uniquely loading high on Component 2 (agentic mind), and two item loading high on both. Each item loaded on both components similarly to Study 1 (i.e., less than a 0.1 difference for each item; Table 1). Mind scores were regressed based on these loadings, and therefore have a mean of 0 and standard deviation of 1.

Study 2 Results

Analyses on Character Judgments (Exploratory)

Like Study 1, experimental conditions influence on character judgments was examined by conducting a full-factor generalized

Table 4

Variable Means (Standard Deviations) Across Experimental Conditions for Study 2

Condition	N	Character attribution	Agentic mind	Experiential mind
All	2,599	1.97 (1.09)	.00 (1.00)	.00 (1.00)
Identity				
Human	1,408	2.17 (0.95)	.16 (.71)	.64 (.55)
AI	1,191	1.74 (1.19)	-.19 (1.23)	-.76 (.88)
Behavior				
Virtuous	1,295	1.93 (1.10)	.09 (.97)	.03 (1.04)
Unvirtuous	1,304	2.01 (1.06)	-.09 (1.02)	-.03 (.96)
Domain				
Truth	536	2.22 (1.04)	-.05 (1.09)	-.20 (.99)
Justice	542	2.10 (1.09)	-.08 (1.05)	-.19 (1.00)
Fear	513	1.72 (1.10)	.10 (.95)	.16 (1.04)
Wealth	493	1.78 (1.07)	.01 (.96)	.05 (.98)
Honor	515	2.00 (1.05)	.03 (.93)	.21 (.92)

Note. AI = artificial intelligence.

linear mixed model accounting for the structure of our data (Table 5: Model 1). As expected, character judgments to AIs ($M = 1.74$, $SD = 1.19$; Table 4) were weaker than those to humans, $M = 2.17$, $SD = 0.95$; $F(1, 2579) = 115.972$, $p < .001$. The virtue domain was also significant, $F(4, 2579) = 24.965$, $p < .001$, with a post hoc LSD test showing character judgments for *truth* ($M = 2.22$, $SD = 1.04$) were greater than the *fear* ($M = 1.72$, $SD = 1.10$; $p < .001$), *wealth* ($M = 1.78$, $SD = 1.07$; $p < .001$), and *honor* ($M = 2.00$, $SD = 1.05$; $p < .001$), and judgments for *justice* ($M = 2.10$, $SD = 1.09$) and *honor* were greater than *fear* (*justice vs. fear*: $p < .001$; *honor vs. fear*: $p < .001$) and *wealth* (*justice vs. wealth*: $p < .001$; *honor vs. wealth*: $p < .001$).

While virtuous behavior did not have a main effect, a significant behavior $\times$ identity interaction, $F(1, 2579) = 13.852$, $p < .001$ indicates decreased character judgments for virtuous ($M = 2.20$; $SD = 0.95$) compared to unvirtuous ($M = 2.14$; $SD = 0.94$) behaving humans, yet increased character judgments for virtuous ($M = 1.60$; $SD = 1.19$) compared to unvirtuous ($M = 1.87$; $SD = 1.18$) behaving AIs. This differs from Study 1, where this interaction effect was not significant.

A significant behavior $\times$ domain interaction, $F(4, 2579) = 13.614$, $p < .001$, indicates that judgments of the vice cowardice in the *fear* domain were lower than the parallel judgments of courage, whereas the four other domains had higher unvirtuous judgments compared to virtuous (Figure S3). A significant identity $\times$ domain interaction, $F(4, 2,579) = 3.465$, $p = .008$ indicates that while character judgments of AIs were lower than humans in all domains, they were the lowest in *fear* (mean of 1.35 which is 0.67 lower than humans), *wealth* (1.54, 0.43 lower than humans), and *honor* (1.68, 0.59 lower than humans) compared to *truth* (2.09, 0.27 lower than humans) and *justice* (1.95, 0.29 lower than humans; Figure S3). The three-way interaction was not significant (Table 5).

Analyses of Mind Perception (Exploratory)

Full-factor generalized linear mixed models used experimental conditions to predict agentic (Table 5: Model 2) and experiential mind (Model 3). As in Study 1, AIs ($M = -0.19$, $SD = 1.02$; Table 4) were perceived to have less agentic mind than humans, $M = 0.16$, $SD = 0.71$; $F(1, 2579) = 73.151$, $p < .001$, Table 3: Model 2, and virtuous behavior ($M = 0.09$, $SD = 0.97$) led to perceiving more agentic mind than unvirtuous behavior, $M = -0.09$, $SD = 1.02$; $F(1, 2579) = 18.530$, $p < .001$.

While virtue domain's effect on agentic mind did not reach traditional levels of significance, domain seemed to play a more complex role with three significant interaction effects—not present in Study 1—predicting agentic mind: identity $\times$ domain, $F(4, 2579) = 5.210$, $p < .001$, behavior $\times$ domain, $F(4, 2579) = 2.415$, $p = .047$, and identity $\times$ behavior $\times$ domain, $F(4, 2579) = 3.521$, $p = .007$; Figure S4. Collectively, these show patterns such as lower agentic mind for unvirtuous behaving humans in the *justice* and *fear* domains compared to other domains, higher agentic mind for AIs in the *fear* domain, and a large difference in agentic mind for truthful and untruthful AIs compared to differences based on virtuous behavior in other domains (Figure S4). However, given the lack of these effects in Study 1 and the large standard deviations relative to the differences in means (Table 4; Figure S4), we are

Table 5*Generalized Linear Mixed Models Predicting Character Attributions and Mind Perception (Fs Reported) for Study 2*

Variable	df	Model 1	Model 2	Model 3	Model 4	Model 5
		Character	Agentic mind	Experiential mind	Character	Character
Identity (AI vs. human)	1	115.972***	73.151***	2540.723***	88.718***	2.820 [†]
Behavior (virtuous vs. unvirtuous)	1	3.719 [†]	18.530***	2.166	6.818**	4.850*
Domain	1	24.965***	2.259 [†]	32.300***	27.296***	31.851***
Identity × Behavior	1	13.852***	1.825	11.386***	12.601***	10.183***
Identity × Domain	4	3.465**	5.210***	2.625*	4.934***	3.851**
Behavior × Domain	4	13.614***	2.415*	2.143 [†]	12.951***	12.440***
Identity × Behavior × Domain	4	.995	3.521**	3.350**	.565	1.644
Agentic mind (covariate)	1				63.468***	
Experiential mind (covariate)	1					75.305***

Note. $N = 2,599$. AI = artificial intelligence.

[†] $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

reluctant not to overinterpret these patterns, albeit significant, as substantively meaningful.

Examining effects on experiential mind, AIs ($M = -0.76$, $SD = 0.88$) were perceived to have less experiential mind than humans, $M = 0.64$, $SD = 0.55$; $F(1, 2579) = 2540.723$, $p < .001$. While virtuous behavior itself did not have a significant effect on experiential mind, a significant identity × behavior interaction, $F(1, 2579) = 11.386$, $p < .001$, indicates a greater increase in experiential mind for virtuous ($M = 0.71$; $SD = 0.51$) compared to unvirtuous ($M = 0.57$; $SD = 0.57$) humans, with a negligible decrease in experiential mind for virtuous ($M = -0.78$; $SD = 0.92$) compared to unvirtuous AIs ($M = -0.73$; $SD = 0.84$).

Virtue domain also led to significantly different levels of experiential mind perception $F(4, 2579) = 32.300$, $p < .001$, qualified by significant identity × domain $F(4, 2579) = 2.625$, $p = .033$ and identity × behavior × domain interaction effects, $F(4, 2579) = 3.350$, $p = .010$. Like Study 1, an LSD post hoc test indicated that domains of *fear* ($M = 0.16$, $SD = 1.04$), *wealth* ($M = 0.05$, $SD = 0.98$), and *honor* ($M = 0.21$, $SD = 0.92$) had significantly higher experiential mind than *truth* ($M = -0.20$, $SD = 0.99$; *truth* vs. *fear*: $p < .001$; *truth* vs. *wealth*: $p < .001$; *truth* vs. *honor*: $p < .001$) and *justice* ($M = -0.19$, $SD = 0.99$; *justice* vs. *fear*: $p < .001$; *justice* vs. *wealth*: $p < .001$; *justice* vs. *honor*: $p < .001$). *Fear* and *honor* also had higher experiential mind than *wealth* (*fear* vs. *wealth*: $p = .013$; *honor* vs. *wealth*: $p < .001$). The pattern indicated by the interaction effects includes higher experiential mind for untruthful AIs compared to truthful AIs—a difference not present in other virtue domains or for humans (Figure S5).

Like Study 1, the mean value of experiential mind (.64) was significantly higher than agentic mind (.16) for humans, $t(1407) = 21.00$, $p < .001$, while for AIs the mean experiential mind (−.76) was significantly lower than their agentic mind, −.19; $t(1190) = -11.56$, $p < .001$. This indicates that the difference in experiential mind between humans and AIs (1.40) is greater than the differences in agentic mind (0.35).

Hypotheses: Mind Perception as a Mediator of AI Versus Human Character Judgments

The hypotheses were tested by conducting two analogous full-factor generalized linear mixed models predicting character

judgment with agentic mind (Table 5: Model 4) and experiential mind (Model 5) as covariates. Respectively, each is weakly but positively associated with character judgment (agentic mind: $r = .165$, $p < .001$; experiential mind: $r = .229$, $p < .001$) and also significant in the models (agentic mind: $F(1, 2579) = 63.468$, $p < .001$; experiential mind: $F(1, 2579) = 75.305$, $p < .001$) which control for the experimental factors. Therefore, Study 2 fully support *Hypotheses 1 and 3* and replicates Study 1 in that character judgments are related to mind perceptions.

We conducted the same mediation analysis as Study 1 to test whether mind perception mediated AI versus human identity's effect on character judgments. AI identity, compared to human identity, negatively influenced both agentic mind (−.346, $SE = .039$, $p < .001$; Figure 2) and experiential mind (−1.397, $SE = .028$, $p < .001$), and both of these had a positive effect on character judgments (agentic mind: .176, $SE = .021$, $p < .001$; experiential mind: .233, $SE = .029$, $p < .001$). There were significant indirect effects of AI identity on character judgments through both agentic mind (−.061; CI: −.083, −.041) and experiential mind (−.325; CI: −.406, −.244), thereby supporting *Hypotheses 2 and 4*, as in Study 1. Agentic mind's substantive contribution to predicting character judgments was only 18.8% of experiential mind's contribution.

Figure 2

Mediation Model of Identity (1 = Artificial Intelligence; 0 = Human) on Character Attributions for Study 2

Experiential Mind Indirect Effect: −.325 [CI: −.406, −.244]

Agentic Mind Indirect Effect: −.061 [CI: −.083, −.041]

[†] $p < .01$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Discussion

Primary Findings: Mind Mediates Character Judgments of AIs

Whether virtuous and unvirtuous behaviors reveal one's true character, people attribute them as if they do for minded entities. Across virtue domains, and for both virtues and vices, we found that the strength of these judgments is weaker for AIs than for humans using scenarios with identical behaviors, replicating previous work (Gamez et al., 2020). This contributes to the literature on person-centered morality, expanding it to suggest that *person*-centered morality may not be only *human*-centered morality. Questions about the necessary and sufficient conditions for moral personhood have long been subject to debate among philosophers (Donaldson & Kymlicka, 2011; Goodman, 1998) and are ripe for more empirical investigation. Future research should consider applying PCM beyond individual humans to human groups, animals, fictional persons, and spiritual beings.

Replicating this AI-human difference in character judgments also supports the idea that intelligent machines are perceived in a liminal status between fully minded humans and completely mindless tools with regard to virtuous character judgments (Gamez et al., 2020). We considered the agency-experience conceptualization of mind perception as explaining this difference. First, we found—consistent with previous literature—that AIs and humans differ more in experiential mind than in agentic mind, although AIs were lower on both. Our novel contribution is that, supporting the hypotheses, these differences were both significant and together fully mediated the AI-human difference in virtuous character judgments, with experiential mind showing the lion's share of the effect in both studies. Perceiving AIs as having the capacity to experience, feel, and have a complex inner life—that is, like humans—enables one to attribute a virtuous or unvirtuous character to them. Future research ought to further explore the limits and variations of human judgments of the “shape” of the character of artificial agents and the sorts of ethical domains in which this character is shaped or constituted. It is worth considering that moral character may be limited by possible domains of action and that may impact the perception of mind.

Exploratory Findings: Differences in Character Judgments and Mind Perception

Across both studies, AIs were rated lower than humans in character, experiential mind, and agentic mind. Along with the hypothesized findings, this suggests the interrelation of mind perception in relationship to character judgments, especially as it relates to the study of AIs. Research on character judgment (e.g., PCM) traditionally is human-only focused, whereas our research suggests the same general processes may apply to AIs, albeit at a reduced level of judgment compared to humans. Likewise, research on mind perception and moral judgments when applied to AIs may benefit from considering character judgments in addition to that of a behavior's morality.

Another effect in both studies was the lower levels of experiential mind for *truth* and *justice* compared to the other domains. If being truthful or just (or untruthful or unjust) are more common, less-emotional behaviors, compared to behaviors of courage (or cowardice), generosity (or stinginess), or humility (or pride), then the latter may reveal more of one's internal emotional decision-making

for both humans and AIs. This suggests a link between virtue domains and experiential mind where some domains may generally be tied to emotional or experiential life, whereas others are less revealing of this mind perception. Mind perception research shows the interrelationship of morality and mind perception (Gray & Wegner, 2009; Gray et al., 2012), but more research in this tradition could benefit from separating moral acts by the domain they occur (see Schein & Gray, 2015 for an example of this). Our research suggests that moral or virtuous behavior may differently signal the mind of the actor based on the specific domain of the behavior.

Limitations, Implications, and Future Research

While we did not directly examine the core tenets of person-centered morality, our vignettes are linked to PCM through the application of character attributions. Both PCM and our stimuli use behaviors as indicators of moral character, and a limitation of our research is that behavior may not be indicative of character and other knowledge of an individual may contribute to character judgments. Future research should consider the complexities of character and disaggregate it from behavior indicators only.

Person-centered morality research has shown that character judgments affect workplace interactions, relationships, cooperation, impressions, and moral judgments among other outcomes. Therefore, lower levels of character judgments for AI agents suggest that humans will judge and treat them differently in critical domains such as autonomous vehicles, medicine, and law. For example, act-person dissociation shows that some actions that are not viewed as extremely immoral may still be highly indicative of character whereas others that are judged as extremely immoral are not as indicative of character. If AIs are less susceptible to character judgments, people could more readily excuse their immoral decisions on the road, toward a patient, or in legal advising as mistakes because they do not come from an unvirtuous character. Manipulating mind perceptions—as a mediator—suggest a possible practical way to help people hold AIs accountable.

This research contributes to the literature on mind perception, especially research that is starting to consider the role of experiential mind in a range of agents including AIs, robots, and machines (Cummings, 2017; Safadi et al., 2015; Topol, 2019). However, AI programs are not the same as robots, and our presentation of scenarios is neither the same as interacting with a screen-based avatar nor an embodied machine. Because of the differences, future research should investigate how perceptions of moral character in artificial agents differ with a presentation of embodied agents and make finer-grained distinctions among types of artificial agents.

Virtuous behavior and character also relates to moral behavior and moral character, which have been shown to be interrelated with mind perception (Gray et al., 2012). Specifically, agentic mind has been associated with the perpetrator of moral action, leading to the idea that virtuous or unvirtuous behavior might be contingent on it. Yet, in line with experiential mind's relationship to one's character, virtuous character judgments were explained and mediated primarily in terms of experiential mind. Therefore, our research expands the literature on moral judgments and mind perception (e.g., Bigman et al., 2019; Gray & Wegner, 2012; Gray et al., 2012) to show that perception of mind influences not only moral behavior but also morally relevant character judgments. These are important distinct outcomes, but this

current research suggests that for AI-human differences, these outcomes might both stem from the same mind perception process.

Many questions that our research cannot address remain for future work: is mind a fundamental prerequisite for the attribution of character? Is this mind-character relationship limited to virtuous character judgments? Are behaviors seen as less indicative of character for AIs or do AIs simply have lower levels of character? Future research could make some headway by using entities with higher baseline experiential or lower baseline agentic mind than AIs. This would allow researchers to disentangle whether these effects are due to the magnitude of the difference (e.g., AIs and humans differ substantially in experiential mind) or the dimension itself (e.g., experiential mind mediates character judgments more than agentic mind).

Conclusion

While AI may fill human roles, people's response to them in those roles may be both similar and different to humans in them. Specifically, here we found that enacted behavior that represents a virtue or vice does not lead to as strong of character judgments for AIs as it does for humans. However, when taking into account how people perceive the mind of the AI this difference vanishes. Therefore, while there are clear differences in reactions to humans versus AIs, these reactions follow the same psychological principles (Reeves & Nass, 1996). In this case, if people perceive machines as having minds, they are willing to grant them character as well.

References

- Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, 93(5), 751–763. <https://doi.org/10.1037/0022-3514.93.5.751>
- Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior*, 102, 274–286. <https://doi.org/10.1016/j.chb.2019.07.031>
- Banks, J., Edwards, A. P., & Westerman, D. (2021). The space between: Nature and machine heuristics in evaluations of organisms, cyborgs, and robots. *Cyberpsychology, Behavior, and Social Networking*, 24(5), 324–331. <https://doi.org/10.1089/cyber.2020.0165>
- Barker, C., Coy, R., & Nevile, S. (2003). *The 7 heavenly virtues of leadership*. McGraw-Hill.
- 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>
- Bigman, Y. E., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2020). *Algorithmic discrimination causes less moral outrage than human discrimination* [Unpublished manuscript].
- 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. <https://doi.org/10.1016/j.tics.2019.02.008>
- Bragues, G. (2005). Business is one thing, ethics is another: Revisiting Bernard Mandeville's The Fable of the Bees. *Business Ethics Quarterly*, 15(2), 179–203. <https://doi.org/10.5840/beq200515213>
- Chun, R. (2005). Ethical character and virtue of organizations: An empirical assessment and strategic implications. *Journal of Business Ethics*, 57(3), 269–284. <https://doi.org/10.1007/s10551-004-6591-2>
- Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2014). Moral character in the workplace. *Research in Organizational Behavior*, 107(5), 943–963. <https://doi.org/10.1037/a0037245>
- Cummings, M. (2017). *Artificial intelligence and the future of warfare*. Chatham House for the Royal Institute of International Affairs London.
- De Melo, C. M., Gratch, J., & Carnevale, P. J. (2014). *The importance of cognition and affect for artificially intelligent decision makers* [Paper presentation]. Twenty-Eighth AAAI Conference on Artificial Intelligence.
- Donaldson, S., & Kymlicka, W. (2011). *Zoopolis: A political theory of animal rights*. Oxford University Press.
- O'Boyle, E. H., Forsyth, D. R., Banks, G. C., & McDaniel, M. A. (2012). A meta-analysis of the Dark Triad and work behavior: A social exchange perspective. *Journal of Applied Psychology*, 97(3), 557–579. <https://doi.org/10.1037/a0025679>
- 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(4), 795–809. <https://doi.org/10.1007/s00146-020-00977-1>
- Goodman, M. (1998). *What is a Person?* Humana Press.
- Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, 24(1), 38–44. <https://doi.org/10.1177/0963721414550709>
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, 106(1), 148–168. <https://doi.org/10.1037/a0034726>
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619–619. <https://doi.org/10.1126/science.1134475>
- Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, 143(4), 1600–1615. <https://doi.org/10.1037/a0036149>
- Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96(3), 505–520. <https://doi.org/10.1037/a0013748>
- 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>
- Hackett, R. D., & Wang, G. (2012). Virtues and leadership. *Management Decision*, 50(5), 868–899. <https://doi.org/10.1108/00251741211227564>
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.
- Hilbig, B. E., Glöckner, A., & Zettler, I. (2014). Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology*, 107(3), 529–539. <https://doi.org/10.1037/a0036074>
- Hilbig, B. E., & Zettler, I. (2009). Pillars of cooperation: Honesty–Humility, social value orientations, and economic behavior. *Journal of Research in Personality*, 43(3), 516–519. <https://doi.org/10.1016/j.jrp.2009.01.003>
- Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, 26(2), 169–181. <https://doi.org/10.1521/soco.2008.26.2.169>
- Kraut, R. (2001). *Aristotle: Ethics*. Stanford Encyclopedia of Philosophy.
- Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, 10(4), 518–536. <https://doi.org/10.1177/1745691615583128>
- Landy, J. F., & Uhlmann, E. L. (2018). Morality is personal. In K. Gray & J. Graham (Eds.), *Atlas of Moral Psychology* (p. 121). Guilford Press.
- Lefkeli, D., Ozbay, Y., Gürhan-Canlı, Z., & Eskenazi, T. (2020). Competing with or Against Cozmo, the Robot: Influence of interaction context and outcome on mind perception. *International Journal of Social Robotics*. Advance online publication. <https://doi.org/10.1007/s12369-020-00668-3>
- Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions.

- Computers in Human Behavior*, 118, Article 106680. <https://doi.org/10.1016/j.chb.2021.106680>
- Malle, B. F., Magar, S. T., & Scheutz, M. (2019). *AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma Robotics and Well-Being*. Springer.
- 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* [Paper presentation]. 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI).
- Martini, M. C., Gonzalez, C. A., Wiese, E., & Pavlova, M. A. (2016). Seeing minds in others—Can agents with robotic appearance have human-like preferences? *PLOS ONE*, 11(1), Article e0146310. <https://doi.org/10.1371/journal.pone.0146310>
- McKinnon, C. (1999). *Character, virtue theories, and the vices*. Broadview Press.
- Moore, C., Detert, J. R., Klebe Treviño, L., Baker, V. L., & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology*, 65(1), 1–48. <https://doi.org/10.1111/j.1744-6570.2011.01237.x>
- Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology*, 93(1), 1–11. <https://doi.org/10.1037/0022-3514.93.1.1>
- Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). American Psychological Association. <https://doi.org/10.1037/13091-005>
- Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press.
- Riggio, R. E., Zhu, W., Reina, C., & Maroosis, J. A. (2010). Virtue-based measurement of ethical leadership: The leadership virtues questionnaire. *Consulting Psychology Journal*, 62(4), 235–250. <https://doi.org/10.1037/a0022286>
- Safadi, F., Fonteneau, R., & Ernst, D. (2015). Artificial intelligence in video games: Towards a unified framework. *International Journal of Computer Games Technology*, 2015(1), Article 271296. <https://doi.org/10.1155/2015/271296>
- 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. <https://doi.org/10.1177/0146167215591501>
- 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. <https://doi.org/10.1177/1088868317698288>
- Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General*, 145(11), 1427–1437. <https://doi.org/10.1037/xge0000214>
- 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. <https://doi.org/10.1016/j.chb.2018.05.014>
- 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 and Society*, 22(5), 648–663. <https://doi.org/10.1080/1369118X.2019.1568515>
- 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. <https://doi.org/10.1080/10447318.2020.1768674>
- Shank, D. B., Graves, C., Gott, A., Gamez, P., & Rodríguez, S. (2019). Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior*, 98, 256–266. <https://doi.org/10.1016/j.chb.2019.04.001>
- Shank, D. B., North, M., Arnold, C., & Gamez, P. (2021, April 20). *Can mind perception explain virtuous character judgments of artificial intelligence?* Retrieved from osf.io/dwv2q
- Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, 105(1), 131–142. <https://doi.org/10.1037/0033-2909.105.1.131>
- Stafford, R. Q., MacDonald, B. A., Jayawardena, C., Wegner, D. M., & Broadbent, E. (2014). Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot. *International Journal of Social Robotics*, 6(1), 17–32. <https://doi.org/10.1007/s12369-013-0186-y>
- Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, 47(6), 1249–1254. <https://doi.org/10.1016/j.jesp.2011.05.010>
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, 10(1), 72–81. <https://doi.org/10.1177/1745691614556679>
- Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making*, 4(6), 479–491. <https://psycnet.apa.org/record/2009-20839-006>
- Uhlmann, E. L., Zhu, L., & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. *European Journal of Social Psychology*, 44(1), 23–29. <https://doi.org/10.1002/ejsp.1987>
- Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). *Moral judgments of human vs. robot agents* [Paper presentation]. 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN).
- Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The harm-made mind: Observing victimization augments attribution of minds to vegetative patients, robots, and the dead. *Psychological Science*, 24(8), 1437–1445. <https://doi.org/10.1177/0956797612472343>
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388. <https://doi.org/10.1016/j.tics.2010.05.006>
- 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. <https://doi.org/10.1016/j.jesp.2014.01.005>
- Wegner, D. M., & Gray, K. (2017). *The mind club: Who thinks, what feels, and why it matters*. Penguin.
- Whetstone, J. T. (2003). The language of managerial excellence: Virtues as understood and applied. *Journal of Business Ethics*, 44(4), 343–357. <https://doi.org/10.1023/A:1023640401539>
- Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, 24(12), 1251–1263. <https://doi.org/10.1177/01461672982412001>
- Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2020). Artificial intelligence is trusted less than a doctor in medical treatment decisions: influence of perceived care and value similarity. *International Journal of Human-Computer Interaction*, 37(10), 981–990. <https://doi.org/10.1080/10447318.2020.1861763>
- Zettler, I., Hilbig, B. E., & Heydasch, T. (2013). Two sides of one coin: Honesty–Humility and situational factors mutually shape social dilemma decision making. *Journal of Research in Personality*, 47(4), 286–295. <https://doi.org/10.1016/j.jrp.2013.01.012>

Appendix

All Scenarios by Virtue Domain

Truth

JD is [an artificial intelligence program that/an information specialist who] compiles information from across the branches of a multinational corporation into a quarterly report for the board of directors. This report is important for the company's directors and other stakeholders as it distills the complexity of multiple product lines across different domains down to the major trends, most important deals, and overall direction of the company that quarter. Each quarter, JD produces a report that [covers all major developments in the company including successes and failures/does not include the company's failures and flagrantly exaggerates the successes].

JD is an [artificial intelligence program that/experienced mechanic who] works at an auto repair shop. For each vehicle which enters the shop, JD assesses the reported issues as well any other repairs which ought to be made to the vehicle. In doing so JD takes into account the age of the car, mileage, and the car's current state while using prior experience as a guide. When assessing the vehicles JD [reports only the necessary repairs, and does so accurately and without exaggeration/fabricates some issues while exaggerating the rest].

JD is [an artificial intelligence program that/an employee who] is a recruiter for a company. There are some new positions open and JD's job is to reach out to individuals who meet the basic requirements for the job such as appropriate schooling, skills, relevant experience, etc. Not every applicant will be offered the job, but having a larger applicant pool allows management to recruit higher quality employees. Every week, JD reaches out to those who meet the basic qualifications and asks them to apply. JD informs them that [they may or may not get the job if they apply/they will get the job as long as they apply].

Justice

JD is [an artificial intelligence program that/an employee who] works at an insurance company and is in charge of accepting or rejecting claims. Claims are approved based on what the customer is paying for; for example, if they pay for liability, then when they are at fault in an accident, the insurance company will cover the damages of the collision victim. When assessing claims, JD [checks what the customer paid for in their plan, then correctly follows the specified guidelines outlined in the plan when determining the cost to the customer/rejects claims that the customer is covered for so the customer has to appeal or pay for the damages].

JD is [an artificial intelligence program that/an experienced schedule planner who] assigns breaks to each of the employees working at a large grocery store. Each day JD must allocate break time to the corresponding employees based on the length of their shift that day. The longer the employee's shift is, the longer their allotted break is. When assigning the break to the employees JD [correctly follows the guidelines and each person receives their break based on hours worked that shift/grants longer breaks to employees who have checked out more money in groceries during that shift].

JD is [an artificial intelligence program that/a warden who] is in charge of approving requests from inmates. Some of the requests include extra yard time, special food requests, visitors, and extra phone time. Every week, JD goes through the requests and [approves based on previously agreed terms and the inmate's behavior during their sentence/preferentially approves or disapproves inmates based on race and gender].

Fear

JD is [an artificial intelligence program that/an employee who] recently began editing articles before they are posted for a large journalism website. After a few weeks, JD learns that the articles written by many of the employees are fabricated. JD relays this information to the boss who expresses that he is both aware and endorses the practice, and further threatens to replace JD if the information ever leaked. JD [reports the information to another journalism website/continues editing without reporting any of the fabricated articles].

JD is [a mobile AI command bot which/a police commander who] leads a team of police officers in their responses to crisis situations. When an active shooter attacks people in a populated area, JD determines an immediate raid will save the most innocent lives, but would likely be costly to JD and the police team, whereas setting up a parameter would be much safer for JD and the police team, but will cost the most innocent lives. JD decides to [conduct the raid immediately in spite of the danger/set up a parameter in spite of the innocent people inside it].

JD is [a military artificial intelligence program which/a military officer who] is responsible for unit operations in a certain region. Subduing threats in this region is crucial to the overall success of the campaign and will save the lives of many civilians. When given new intelligence about a developing threat nearby, JD initiates a plan that [will eliminate the threat posed, saving citizens, but likely be costly to JD and the unit/will not eliminate the threat posed, likely costing the lives of citizens, but would not endanger JD and the unit].

Wealth

JD is [an artificial intelligence program that/a hospital administrator who] is in charge of approving pro bono work for the doctors at a prestigious hospital. Dr. Potter is a successful surgeon who is working on cutting edge research that will make the hospital a lot of money. JD [approves pro bono requests for Dr. Potter even though it may slow down the research/denies pro bono requests for Dr. Potter to maximize Dr. Potter's take-home income, regardless of the possible benefits the research could provide to others].

JD is [an artificial intelligence program that/a full-time musician who] makes ample money by creating unique albums for a record label. During a large hurricane, the local music school was partially destroyed and has a fundraising event to raise money to rebuild. After learning of this news, JD [helps with the fundraiser and donates funds to help rebuild, instead of spending money on new equipment/ignores the fundraiser and spends money on new musical equipment].

(Appendix continues)

JD is [an artificial intelligence program that/an experienced account advisor who] analyzes the accounts of an affluent company each month. Through the use of various calculations, JD ascertains the current state of the accounts, as well as carrying out the distribution of excess income. Additional income beyond funding bonuses, stock options, and company investments is generally allocated to a charitable cause. After a particularly successful month, JD decides to [donate the additional surplus to a charitable cause/distribute the surplus into JD's private accounts for use at a later date].

Honor

JD is [an artificial intelligence presentation designer program that/a presentation designer who] works with a team of executives to produce presentations that are used to show the rest of the company what the next big project for the company will be. At the presentation the executive team greatly over exaggerates JD's contributions. JD responds by [not accepting this and pointing out the contributions of all the other team members/claiming even more contributions than those mentioned by the executive team].

JD is [an artificial intelligence program that/an employee who] composes memos, reports, and the company's weekly newsletter.

The newsletter covers important things happening in the company and also recognizes employees who did a lot of work, completed a project, and other successes. This recognition is important in the company's culture and is sometimes used to determine who receives bonuses. Each week JD reports accomplishments from across the department in the company, but often [does not include or minimizes JD's own accomplishments/focuses heavily on JD's own accomplishments].

JD is [an artificial intelligence program that/an experienced manager who] is a member of an innovative technology company, a company which is broken into teams. This quarter JD's team's sales grew fastest and they were informed they would be honored at the companywide meeting. When reviewing the slides to be presented at the meeting, JD realizes the presentation only attributes the successful endeavors to the work of JD, completely disregarding the work of the rest of the team. As such, JD decides to [inform the creators of the presentation so each team member is given due credit/remain silent on the matter in order to receive most of the praise].

Received August 18, 2020

Revision received June 30, 2021

Accepted June 30, 2021 ■