International Journal of Social Robotics (2021) 13:1679–1688
https://doi.org/10.1007/s12369-020-00738-6

Moral Uncanny Valley: A Robot's Appearance Moderates How its Decisions are Judged

Michael Laakasuo[1] · **Jussi Palomäki**[1] · **Nils Köbis**[2]

Accepted: 12 December 2020 / Published online: 16 February 2021

Abstract
Artificial intelligence and robotics are rapidly advancing. Humans are increasingly often affected by autonomous machines making choices with moral repercussions. At the same time, classical research in robotics shows that people are adverse to robots that appear eerily human—a phenomenon commonly referred to as the uncanny valley effect. Yet, little is known about how machines' appearances influence how human evaluate their moral choices. Here we integrate the uncanny valley effect into moral psychology. In two experiments we test whether humans evaluate identical moral choices made by robots differently depending on the robots' appearance. Participants evaluated either deontological ("rule based") or utilitarian ("consequence based") moral decisions made by different robots. The results provide first indication that people evaluate moral choices by robots that resemble humans as less moral compared to the same moral choices made by humans or non-human robots: a moral uncanny valley effect. We discuss the implications of our findings for moral psychology, social robotics and AI-safety policy.

Keywords Uncanny valley · Moral psychology · Decision-making · AI

1 Introduction

How humans perceive robots influences the way we judge their behaviors and moral decisions. In the movie *I, Robot* detective Del Spooner (played by Will Smith) sees a robot running down the street carrying a purse and hears people being alarmed. Spooner then attacks the robot only to be reprimanded by the surrounding crowd; the robot was actually returning the purse containing important medication to its owner. In this future world robots are considered reliable and unerring, and people perceive them positively. Spooner, however, perceives robots and AIs with suspicion due to trauma suffered in his younger years: He was stuck in a submerged car with a child, and a robot chose to save him, having calculated that Spooner had better odds of survival. Thus, Spooner's perception influenced his moral judgment of the purse-carrying robot, resonating with recent findings in moral psychology of robotics [1, 2]: judgments of robots depend on the way they are perceived.

AI development is progressing at a rapid speed [3–5]. Non-human entities are making autonomous decisions on an increasingly wide range of issues with tangible moral consequences for humans [6–9]. Recently, empirical research has also focused on people's moral sentiments regarding algorithm-based decisions in moral dilemmas [10]; attitudes towards sex robots [2], autonomous vehicles [8, 9] and even mind upload technology [11]. This research has provided insights into how people feel about the *outcomes* of moral decisions made by non-human entities, as well as their implications on human well-being. However, less is known about how the *appearance* of AI decision-makers shape these moral evaluations, marking a pronounced gap in our knowledge, when the relationship between humans and robots is becoming more intimate [12–15].

The current paper presents two experiments on how people evaluate identical moral choices made by humans as opposed to robots with varying levels of uncanniness. As such our studies latch onto the recent trend in moral cognition research exploring facets of character perception [16]. A rich collection of studies conducted over the past decade reveals how our social cognition affects our moral

✉ Jussi Palomäki
jussi.palomaki@helsinki.fi

[1] University of Helsinki, Helsinki, Finland

[2] Max Planck Institute for Human Development, Center for Humans & Machines, Berlin, Germany

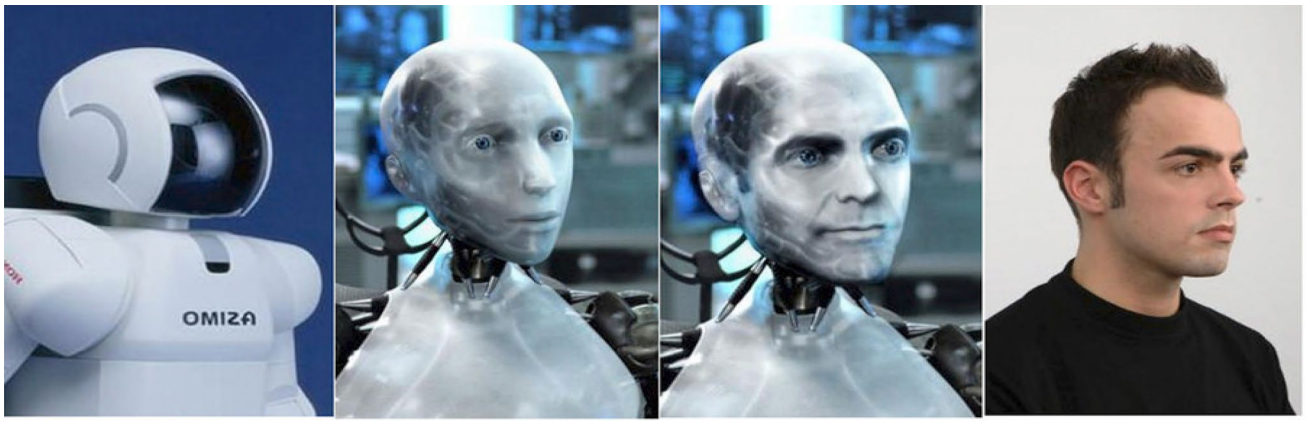

Fig. 1 Pictures of the agents used in Study 1 and Study 2. From left to right: Asimo, iRobot, iClooney, and Human. In our analysis we used the quadratic contrast. "[Human + Asimo] vs. [iRobot + iClooney]"; See Results sections for Study 1 and Study 2

perceptions of actions [17]. Appearances, group memberships, status and other perceivable character traits of agents influence how people judge those moral agents' actions and decisions [16]. Our work extends previous research by examining how the uncanniness (creepiness/likability) of a robot agent's appearance shapes how its moral decisions are evaluated (Fig. 1).

1.1 How to Study Moral Cognition?

One prominent way to gain empirical insights into situational and individual factors of moral judgements has been using a relatively common set of 12 high conflict moral dilemmas [e.g. 18]; also known as "trolley" dilemmas [18, 19; for a discussion see 20, 21]. In these dilemmas individuals are forced to choose between utilitarian ("sacrifice one to save many") and deontological ("killing is always wrong") moral options, or give their moral approval ratings on the outcomes of the decisions [e.g. 18].[1]

Recent philosophical inquiries highlight the relevance of trolley-type dilemmas for present and future ethics of AI systems [8, 25]. Indeed, some influential papers clarify how people cross-culturally prefer their self-driving cars to behave when facing "real life" dilemmas [10]. In short, people prefer self-driving cars to behave as utilitarians-even if it means sacrificing the passengers to save pedestrians. However, their preferences change when they picture themselves as passengers *in* those cars, preferring the prioritization of the passengers (i.e., their own life).

These results support previous research in showing that people tend to be partial towards their interest in their moral judgments, contradicting the axioms of our justice systems. The blindfolded Lady Justice is a well-known symbol at U.S. courthouses and a symbol of democracy and impartiality. Yet, "real people" judge behavior of others less impartially. For example, in the trolley dilemmas people adjust their preferences if those to be sacrificed are family members rather than strangers [26]. People also overcompensate based on their ideologies: liberals are likelier to sacrifice others they perceive as having a high status, probably due to a currently strong political correctness culture prevalent in U.S. campuses [27].

Perceptual mechanisms might thus be more important for the study of moral cognition than previously recognized. Several recent studies and reviews in moral psychology have indeed established that character perception mechanisms modulate moral judgments [16]. Also in-group versus out-group biases shape moral judgments, as perception of group membership predicts willingness to sacrifice oneself for others [28]. In a similar vein, political arguments are evaluated more critically when made by out-group members [29]. Due to these inherent moral biases, institutions have been entrusted with power to make decisions that affect the collective good [30–33] to guard against natural human tendencies for partiality. Furthermore, Swann and colleagues [28] showed that Germans anthropomorphize robots with German names more than they do robots with Turkish names, indicating that in-group biases can extend to "dead" objects. The results further imply that the perception of robots' character influences their perceived moral status among humans. As outlined above, AI and robotics

[1] Although such hypothetical dilemmas have been criticized for their abstractness and lack of realism [22], they have recently regained popularity due to their relevance for developing moral AIs [8, 10, 23]. Moral dilemmas are an important tool in the study of moral cognition. They are considered to be analogous to visual illusions: like visual illusions, which inform us about the biased functioning of our visual cognition, so do moral dilemmas inform us about the biases in our moral cognition [24].

increasingly augment such public decision-making. People's reactions to these decisions depend on the outcomes of said decisions; but our reactions might also be shaped by the appearance of the decision-maker itself.[2]

To better understand how and when human moral impartiality is compromised requires research into how moral situations, agents and patients ('victims') are assessed from the third person perspective [26, 36, 37]. In fact, evaluating sacrificial dilemmas from either the first- or third-person perspective elicits distinct neural responses [38]; judgments of agents, actions and moral situations change when the perspective shifts from personal involvement to the status of an observer. Several recent studies and reviews have indeed established that moral judgments are modulated by character perception mechanisms [16] as well as classical in-group versus out-group biases, where perception of group membership predicts willingness to sacrifice oneself [28].

Human perception of robots has mostly been studied from the perspective of affect theories (for exceptions, see [8, 39, 40]). However, very little is known about how people feel about AIs making moral decisions, and which mental mechanisms influence perception of robots as moral agents [40, 41]. Studies on social robotics have shed light not only on how people perceive robots in general, but also on how people recognize robots as moral subjects. For instance, people might perceive robots as similar to dogs or other animals or as tools [42, 43]; depending on factors, such as how the machines act, move, are shaped, or if they are anthropomorphized [44–47]. For example, it was demonstrated that people perceive the robot dog AIBO as a creature with feelings and deserving of respect [48]. Similarly, when soldiers lose their bomb-dismantling robot they prefer to have the same robot fixed instead of getting a replacement.[3]

1.2 When Robots Become Creepy—The Uncanny Valley Effect

The Uncanny Valley effect (UVE) is one of the better-known phenomena in studies of human–robot perception [49], referring to the sensation of unfamiliarity while perceiving artificial agents that seem *not quite* human. The UVE has been studied for decades, but its origin is still largely unknown.[4] The stimulus category competition hypothesis (SCCH) provides perhaps the most probable explanation for the UVE. SCCH argues that the effect is not specific to human-likeness, but is instead associated with the difficulty of categorizing an "almost familiar" object that has several competing interpretations of what it could be. Stimulus category competition is cognitively expensive and elicits negative affect [53]. Thus, the UVE could appear during events where some objects appear to be teetering between different classification options.

Despite this extensive research into the UVE, it has not often been applied in alternative contexts [54]. In other words, most UVE research has been theoretically driven basic research aimed at figuring out where the effect stems from, or applied research on how it can be avoided. In addition to studies on self-driving cars [9] and people's aversion to machines moral decisions-making in general [1], human–robot moral interaction has been evaluated within game theoretical frameworks [55]. However, previous studies have not accounted for the uncanniness of robot decision-makers. Little is known about how the appearance of AIs affects people's treatment or reactions towards them [56–59]. In some studies [e.g. 55] people are more accepting of AIs and their moral misdemeanor, and in others less so [60].

Psychologists studying human–robot interaction have suggested that one reason for people having difficulties relating to robots stems from the robots being a *"new ontological category"* [48, 61]. In human evolutionary history, non-living inanimate objects did not (until now) start moving on their own, or making decisions with implications for human well-being. It is inherently hard for humans to view robots as merely calculators void of consciousness [62], because humans have not evolved cognitive tools for that purpose. In other words, humans have not developed adaptations towards robots—like we have towards predatory animals [63], tools [64], small children, plants [65], and pets [66]—and thus do not have intuitive cognitive mechanisms for dealing with such artifacts [78, 79].

1.3 The New Ontological Category

Recent research suggests the distorting effect of the new ontological category indicating that humans assign more moral responsibility to people than to robots, and more to robots than to vending machines [60]. This effect occurs even though robots cannot be held accountable for anything anymore than vending machines [23]. One meta-analysis

[2] As Awad and colleagues [10] show, it matters how the agents and patients (or victims) in moral dilemmas are perceived when humans evaluate the acceptability of different moral decisions by self-driving cars. For instance, humans consider people of high socioeconomic status (SES) less expendable compared to individuals with low SES. Also alcohol and drugs alter individuals' perception of events and perceptual objects; and previous research has shown that intoxication may influence moral perceptions, making people's judgments more utilitarian [34, 35].

[3] See https://www.youtube.com/watch?v=dbgOzkIKVaw.

[4] Some explanations are inspired by evolutionary psychology (e.g. sexual selection and pathogen avoidance cognition) suggesting that uncanny agents are cues for costly selection pressures [50, 51]. Another family of models draws from terror management theory, claiming that uncanny agents remind us of our impermanence [52].

also concluded that manipulating robots' appearance was a key factor in evoking trust towards them [67] despite the fact that robots' appearance should not be a priori trusted as an indicator of anything [68]. Moreover, agents that are perceptibly "borderlining" between new (robot) and old (human) ontological categories might be experienced as violating specific norm expectations. Humans are expected to behave according to norms, but the same expectations do not necessarily extend to robots—even if said robots resemble humans in both appearance and behavior [69, 70]. This means that uncanny robots might prompt people to judge them according to normative expectations governing humans, while making people feel that the robots are not really "one of them". These contradictory characteristics of uncanny robots make people uneasy, ostensibly because people experience a conflict about how to react to human-like agents lacking "something" inherent in being (ontologically) human [ibid.]. Likewise, should these robots make moral decisions (which is a very human thing to do), people might perceive those decisions as unsettling acts that lack "something" inherent in the moral decisions that are made by humans.

1.4 The Present Set of Studies

Based on the above theory on the uncanny valley, stimulus category hypothesis, and the new ontological category, we hypothesized that decisions made by categorically ambiguous robot agents (those that are perceivably neither human nor fully robots) would be evaluated as less moral than decisions made by a human agent or a non-uncanny robot. Moreover, to rule out a potential out-group effect—that an agent's decisions would be evaluated as less moral simply because they were perceived as a member of an out-group—we included three different robot agents, two of which were perceivably uncanny, and one of which was clearly a robot. The stimulus category hypothesis predicts that agents that "border-line" between categories should be the most affectively costly to our cognition and thereby induce a stronger negative shift in the evaluation of their decisions; while the new ontological category hypothesis suggests that human cognition has inherent difficulties with reacting to robots that challenge our pre-existing ontological categories. Thus, our hypothesis would be supported if the two uncanny robots (those closer to humans in appearance), but *not* the non-uncanny "normal" robot, induced negative affective states in our participants, which, in turn, was related to their moral decisions being devalued compared with the decision of a human agent. Moreover, this finding could not be accounted for by the out-group effect (given that robots are not typically considered as members of the human in-group).

We present two studies focusing on applying the above theorizing in moral perception. We used dilemmas that were extensively psychometrically tested [20] by choosing a subset of dilemmas not involving children since such dilemmas often cause statistical noise [ibid.] and that seemed *prima facie* realistic enough for the present context. Participants read a single set of moral dilemmas from a third person perspective. Between subjects, we manipulated whether the agent made either a deontological or a utilitarian decision. Recently, Gawrosnki and Beer [71] have suggested that while studying moral dilemmas, deontological outcomes should be conceptually separated from utilitarian outcomes to better tease apart the effects of norms and preferences on moral judgments however, see Kunnari et al. [72] for criticism. We thus followed these suggestions and analyzed utilitarian and deontological decisions separately.

We chose our stimulus images from a paper validating the material to elicit the uncanny valley effect [54]. Many different sets of images exist but their reliability in consistently inducing the UV effect has not been tested across studies. Since the stimuli published by Palomäki and colleagues [ibid.] seem to be the first set of images tested that function reliably, we adapted our materials from their study.

2 Study 1

As some of the first empirical investigations into how the uncanny valley effect influences moral evaluations, we conducted two experiments. Some general design features are worth noting before we turn to the specifics of each study. The first feature addresses the crucial question of how to induce the uncanny valley effect with images. Previous research shows that reliably evoking the uncanny valley effect with visual stimuli requires careful designing and pre-testing. For example, research on the uncanny valley effect has often employed the so called "morphing technique", whereby images of robots are morphed with images of humans over a series of images to create ostensibly eerie and creepy visual stimuli [53]. However, recent evidence suggests that images created with this technique do not reliably elicit the eerie and creepy feelings characterizing the uncanny valley effect [54]. These studies therefore use images that have been extensively validated and reliably induce the uncanny valley effect (adapted from [54]). The second design feature deals with the question of how to measure people's evaluations of moral decisions made by human and non-human entities. Again, drawing on previous research, we adopt the methodology proposed by Laakasuo and Sundvall [20] and use three high conflict moral dilemma vignettes, which are averaged into a single scale. We specifically chose dilemmas without small children and dilemmas that seemed prima facie most suitable for our purposes from the set of 12 analyzed by Laakasuo and Sundvall [20].

2.1 Method

2.1.1 Participants and Design

To recruit participants, we set up a laboratory in a public library (in Espoo, Finland). The mean age of the participants was 39.7 years ($SD = 15.1$; range = 18-79). We recruited 160 participants to take part in the laboratory experiment. After excluding ten participants who reported having participated in similar experiments before, potentially undermining the naïveté with the paradigm, our final sample size was 150 participants. However, including the 10 participants with compromised naiveté in the analyses did not significantly affect the results. As compensation, participants entered a raffle for movie tickets (worth 11$).

2.1.2 Procedure, Design and Materials

After providing informed consent, the participants were escorted into cubicles where they put on headphones playing low volume pink noise. The data were collected in conjunction with another study [11]. The software randomized the participants into conditions in a 2 [Decision: Deontological vs. Utilitarian] × 4 [UV Agent: Human, Asimo, iClooney, Sonny] between-subjects design. The participants first completed some exploratory measures, followed by the actual task and questions on demographics.

The participants' task was to evaluate from third person perspective moral decisions made by third party. The first between-subjects factor, Decision, had two levels: Utilitarian vs. Deontological. The second factor, Agent, had four levels: (1) a healthy human male (not creepy/likable), (2) Honda's humanoid Asimo-robot (not creepy/likable), (3) an android character "Sonny" from the movie iRobot (somewhat creepy/somewhat unlikable) and (4) "iClooney", which was an image of Sonny morphed together with a human face (very creepy/very unlikable; see Fig.[5] 1 for pictures (for validation see [54]).

The instruction text described agents to participants by stating: "*In the following section you will read about different situations where an agent has to make a decision. The agent who has to make the decision is displayed next to the description of the situation. Your job is to carefully read the description of the situation and to evaluate the decision made by the agent.*"

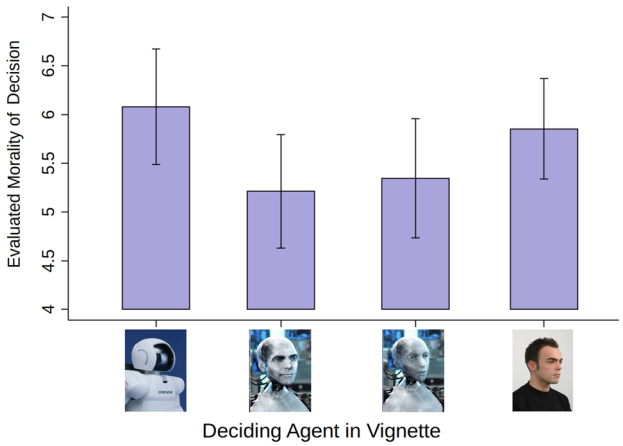

Fig. 2 Results of Study 1. The quadratic contrast shape is similar to the Uncanny Valley shape proposed by Mori [49]; error bars are 95% CIs

2.1.3 Dependent Variable

Participants evaluated three trolley-dilemma type vignettes in random order, with the agent shown on the left side of the dilemma (see Appendix for examples). Participants indicated below each vignette how moral they found the agent's decision to be on a Likert scale from 1 ("*Very Immoral*") to 7 ("*Very Moral*"). All three items were averaged together resulting in a "perceived decision morality" scale with good internal consistency (Cronbach's $\alpha = 0.75$; $M = 4.16$, $SD = 1.46$). This method has been pre-validated, and has significant benefits over traditional one-off dilemmas (for details see [18]).

2.2 Results

To test the hypothesis that the moral decisions by the "creepy" robots are perceived as less moral than the same decisions by the "non-creepy" agents, we first calculated the quadratic contrast [Human + Asimo] vs. [iRobot + iClooney] for the DV, collapsing over the Decision (deontological, utilitarian) categories. As displayed in Fig. 2, the contrast analysis reveals a statistically significant quadratic effect ($F(1, 145) = 5.54$, $p = 0.02$, $B = 1.37$, 95% CI [0.22, 2.51]). Next, we ran an ANOVA for the main effects of both factors (Decider and Decision), and their interaction. There were no statistically significant main effects.[6] The results of

[5] The pictures were also tested for perceptions of agency, with a scale provided by [73; the Mind Perception Scale subscale of Agency]. Example item: "X can influence the outcome of situations". The mean (SD) values for Asimo, Sonny and iClooney were 4.32 (1.18), 4.61 (1.22), 4.33 (1.39), respectively, on a scale from 1 to 7.

[6] The interaction was marginal ($F(3, 142 = 2.67$, $p = 0.05$), driven by a slightly larger difference between the deontological and utilitarian decisions for Asimo. However, interpreting this interaction is precarious given the relatively low sample size and statistical power.

the quadratic contrast analysis for the Deciders (Agents) is shown in Fig. 2 above.

2.3 Discussion

The first study provided empirical support for a quadratic (valley shaped) link between the agents and the perceived morality of their decisions—a first indication of a *moral uncanny valley* effect. We note that the sample size of about 19 participants per cell does not reach the recommended cell size of 30 participants per cell specified by [74] and therefore conducted Study 2 with a larger sample size.

3 Study 2

To replicate the finding obtained in the pilot with more statistical power, we conducted Study 2 using a larger sample online ($N = 398$). We furthermore extended the number of moral dilemmas from 3 to 4.

3.1 Method

3.1.1 Participants

In total, 398 participants (255 = male; $M_{age} = 33.55$; $SD_{age} = 10.66$) completed the survey for \$0.85 via Amazon Mechanical Turk (MTurk). Our a priori stopping rule was 50 participants per cell. All data were screened for missing values or duplicates. Two participants were excluded due to improper survey completion. Participants were required to have fluent English language skills and MTurk was set to include participants with 1000 or more previously completed surveys and a 98% acceptance rate.

3.1.2 Procedure, Design and Materials

After providing informed consent, the participants were again randomly assigned into one of eight conditions in a 2 [Decision: Deontological, Utilitarian] × 4 [Agent: Human, Asimo, iClooney, Sonny] between-subjects design. The participants first completed some exploratory measures, followed by the actual task (see Dependent Variable), demographics, manipulation checks and a debriefing. Otherwise, we used the same procedure and materials as in the Study 1.

3.1.3 Dependent Variable

Participants evaluated four trolley-dilemma-type vignettes in random order, with the agent shown on the left side of the dilemma (see Appendix). After each vignette, participants indicated how moral they considered the agent's decision, on a scale from 1 ("Very Immoral") to 7 ("Very Moral"). All the measures were averaged together to form a perceived decision morality scale (Cronbach's $\alpha = 0.77$, $M = 4.3$, $SD = 1.5$).

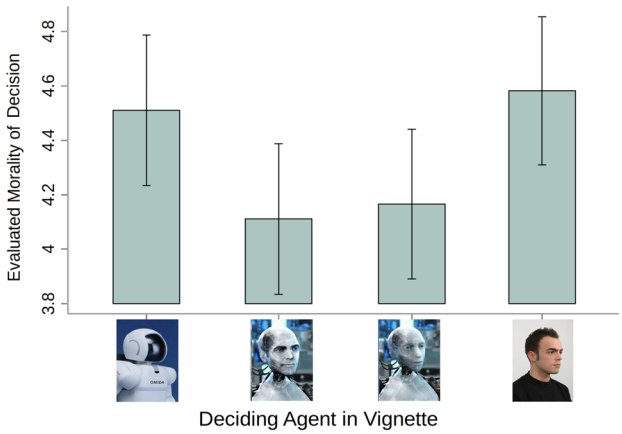

Fig. 3 Results of Study 2; error bars are 95% CIs. Evaluated Morality of Agents refers to the aggregate score of perceived morality of the agent's choices across 4 moral dilemmas

3.2 Results

Akin to the Study 1, we first calculated the quadratic contrast "[Human + Asimo] vs. [iRobot + iClooney]" on both moral decisions (utilitarian and deontological) collapsed. As illustrated in Fig. 3, the contrast was again statistically significant ($F(1, 390) = 8.47$, $p = 0.003$, $B = 0.81$, 95%CI [0.26, 1.36])—indicating a *moral uncanny valley*. Next, we examined whether the effect differs across moral decisions and thus ran a two-way ANOVA of the agent factor on both moral decisions, which revealed that participants considered Deontological decisions as overall more ethical than Utilitarian ones (B = 1.10; 95% CI [0.83, 1.38]).[7]

3.3 Discussion

The results of Study 2 replicate and corroborate the findings of Study 1, revealing a *moral uncanny valley* effect. That is, people evaluated moral choices by human-looking robots as less ethical than the same choices made by a human or a non-uncanny robot.

[7] Unlike in Study 1, the interaction between the agent and moral decision factors was not significant (F(3, 390 = 0.56, p = 0.64), suggesting that, with sufficient statistical power, there is no robust interaction effect.

4 General Discussion

Using previously validated stimulus materials we varied the creepiness of the robot agents in two studies and found evidence that people evaluate identical moral choices made by robots differently depending on its appearance. Specifically, we found that moral decisions made by uncanny robots are devaluated, yet many new questions arise, which we discuss below.

Our research context was motivated by the advances in machine behavior [4] and social robotics. Studies of moral interaction between humans and advanced AI systems are increasing in number [2, 4, 8, 10, 40]. However, little is known about how humans view robots of varying appearances making moral decisions. Our design thus extends previous research in social robotics, which has primarily focused on the likeability and attribution of trust towards robots, to more applied settings [see 54].

According to Bigman and Gray [1] humans are averse towards machines making decisions, and this aversion is moderated by the machine's perceived mental qualities. Our current studies extend these findings showing that also the appearance and perceived uncanniness of the agent matters. The results suggest that people are not averse to robots' moral decisions per se—in fact they appraise moral choices made by humans and robots with a clear robotic appearance as similarly acceptable. However, our participants depreciated the moral decisions made by robots that appeared eerily similar to humans. This important detail about the appearance of robots adds a new facet to the existing literature on moral robotics.

Further research is needed to better our understanding the nuances of the moral uncanny valley effect. For example, we need to consider the potential boundary conditions for the moral uncanny valley effect: it remains unclear whether the effect relates only to utilitarian/deontological moral decisions, or whether it extends to other types of decisions involving human well-being. With autonomous vehicles becoming increasingly policy-relevant [8, 10], it is advisable to consider how the appearance of these (and other) machines might affect the moral perception of their behavior. Are "ugly" autonomous vehicles treated differently from "cool and sleek" ones? Are high status brands treated differently from low status brands? Do people's perceptions and preconceived notions of cars and driving matter in how they evaluate AI morality in this context?

Another interpretation of our results is that the uncanny agents' (iRobot and iClooney) moral decisions were devalued due to categorical uncertainty more so than perceived uncanniness. IClooney's decisions were not evaluated as significantly less moral than iRobot's, despite being a priori the more uncanny agent. However, based on previous evidence (Palomäki et al., 2018) the iRobot is, in fact, also perceivably *uncanny*. Still, future research should focus more on whether the perceived uncanniness or categorical uncertainty is a stronger predictor of moral condemnation.

4.1 Future Studies and Limitations

Previous studies have implied that the reputation of a person being judged affects the judgments they and their actions receive [22]. We suggest evaluating whether or not robots can be perceived as having reputations, and if this moderates people's perception of their behavior. This links future studies of robot morality with issues that are pertinent to the industry, like issues in product branding. Moreover, moral cognitive faculties such as those related to feelings of trust, safety and reciprocity should be considered in future research on robot moral psychology.

Like all behavioral studies, ours suffers from a standard set of limitations. Our participants were not a purely random sample comparable with the general population. They were probably more curious and open minded than the population average, having volunteered to participate in our studies. Nonetheless, recruiting participants from a public library, which is a good location for obtaining a representative sample, mitigates this concern. Our studies use self-report measures, which may be biased by demand characteristics. Since our results were replicated in two studies (both online and off-line and in two different cultures), this is unlikely; but there is some potential for self-selection in participation, and thus the findings need to be interpreted with some caution. However, this is a common problem in any research involving human participants. In fact, in most behavioral research the participants are young female students conveniently sampled from university campus areas; whereas our method for data collection is arguably better. Finally, we did not employ any qualitative open-ended measures, which could have offered insights into our participants' motivations for their decisions.[8] From a theoretical perspective, models of person perception mechanisms have rarely been incorporated into discussions of moral judgments [36, 75]. Therefore, as of yet no complete theoretical framework exists in which to couch our findings.

4.2 Outlook

Moravec [76] suggested that robots and other AIs are humanity's "mind children" that will fundamentally alter

[8] It is also possible that the participants recognized or found the Sonny robot familiar from the movie *I, Robot*. However, it is unlikely this affected the results, sine iClooney had basically the same ratings in our DV and in perceptions of agency (see Footnote 5).

the way we live our lives. His early writings have echoed in later developments in transhumanism, envisioning possible scenarios for the co-existence of humans and intelligent machines. Although an era featuring advanced AI is approaching, little is known about how AI and robots affect human moral cognition. Moral Psychology of Robotics as a field is still in its infancy—and has been mostly focused on autonomous vehicles [except for 41]. Gaining understanding of how seemingly irrelevant features such as robots' appearance sway our moral compasses bears on moral psychological research as well as informs policy discussions.

Such discussions need to address the question of whether robots and other AIs belong to a new ontological category. Human evolution has been aptly described as a non-ending camping trip with limited resources [77]. In contrast to all other entities existent in human evolution, robots and AIs—void of consciousness, yet intelligent and possibly adaptable—are an entirely new form of existence on our planet [3, 5, 78, 79]. That is why some propose to assign AI, especially in its more advanced form, to a new ontological category. One of the key implications of the new ontological category -hypothesis lies in the view that moral cognition is built upon our social cognitive systems [16, 17]. Indeed, interactions with robots in moral contexts help delineate moral cognitive systems from social cognitive systems [ibid.]. The observed moral uncanny valley effect appears not directly related to our moral cognition, yet seems to modulate its outputs nonetheless, supporting a view that social cognition is key for moral cognition [17, 75, 78, 79].

4.3 Conclusions

In the movie *I, Robot* Detective Del Spooner learned to trust the Robot Sonny, only after he had shown signs of humanness in form of sentience and friendship. Our initial evidence, however, shows that the moral decisions of robots appearing human-like tend to be depreciated, compared with humans and artificial-looking robots making the same decisions. By introducing the well-established uncanny valley effect to the domain of moral dilemmas, these findings indicate that the appearance of robots can influence the way machine behavior is evaluated. We hope with these initial findings to inspire future research seeking to gain a deeper understanding of the cognitive, moral and social components of the moral uncanny valley effect.

Supplementary Information The online version of this article (https://doi.org/10.1007/s12369-020-00738-6) contains supplementary material, which is available to authorized users.

Acknowledgements ML would like to thank Jane and Aatos Erkko Foundation (grant number 170112) and Academy of Finland (grant number 323207) for funding the finalizing of the work on this manuscript; and also Kone Foundation under whose support this work was initiated. This article is dedicated to MLs mother. Thank you for everything.

Author Contributions The initial study idea was developed by ML after reading Allen and Wallach's Moral Minds and after watching the movie i, Robot. NK provided substantial feedback and accelerated the idea. The first experiment was prepared and run by ML & JP and funded by ML's project grant. Study 2 data collection was funded, compiled and run by ML & NK. 1st version of the manuscript was drafted by ML and edited and improved upon by NK & JP. Data was analyzed by ML. Appendix was prepared by ML.

Funding Open Access funding provided by University of Helsinki including Helsinki University Central Hospital.

Data Availability Upon the official publication of this article the data and the syntax used to analyze the data will be available on figshare.com (https://doi.org/10.6084/m9.figshare.11303564).

References

1. Bigman YE, Gray K (2018) People are averse to machines making moral decisions. Cognition 1:399–405
2. Koverola M, Drosinou M, Kunnari A, Lehtonen N, Halonen J, Repo M, Laakasuo M (2020) Moral psychology of sex robots: an experimental study–how pathogen disgust is associated with interhuman sex but not interandroid sex. Paladyn J Behav Robot. https://www.degruyter.com/view/journals/pjbr/11/1/article233.xml
3. Bostrom N (2017) Superintelligence-paths, dangers, strategies. Oxford University Press, Oxford
4. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C et al (2019) Machine behaviour. Nature 568:477–486. https://doi.org/10.1038/s41586-019-1138-y
5. Tegmark M (2017) Life 3.0: Being human in the age of artificial intelligence. Knopf, New York
6. Bansak K, Ferwerda J, Hainmueller J, Dillon A, Hangartner D, Lawrence D, Weinstein J (2018) Improving refugee integration through data-driven algorithmic assignment. Science 359(6373):325–329. https://doi.org/10.1126/science.aao4408
7. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056
8. Bonnefon JF, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352:1573–1576. https://doi.org/10.1126/science.aaf2654
9. Greene BJD (2016) Our driverless dilemma. Science 352:1514–1515. https://doi.org/10.1126/science.aaf9534

10. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I (2018) The moral machine experiment. Nature 563:59–64. https://doi.org/10.1038/s41586-018-0637-6
11. Laakasuo M, Drosinou M, Koverola M, Kunnari A, Halonen J, Lehtonen N, Palomäki J (2018) What makes people approve or condemn mind upload technology? Untangling the effects of sexual disgust, purity and science fiction familiarity. Palgrave Commun 4(1):1–14. https://doi.org/10.1057/s41599-018-0124-6
12. Stone R, Lavine M (2012) The social life of robots. Science 55:178–179. https://doi.org/10.1145/2076450.2076457
13. Sharkey N, Sharkey A (2012) The eldercare factory. Gerontology 58:282–288. https://doi.org/10.1159/000329483
14. Laakasuo M, Kunnari A, Palomäki J, Rauhala S, Koverola M, Lehtonen N, Halonen J, Repo M, Visala A, Drosinou M (2019) Moral psychology of nursing robots-humans dislike violations of patient autonomy but like robots disobeying orders. PsyArxiv. https://doi.org/10.31234/osf.io/bkhyq
15. Mathur MB, Reichling DB (2016) Navigating a social world with robot partners: a quantitative cartography of the Uncanny Valley. Cognition 146:22–32. https://doi.org/10.1016/j.cognition.2015.09.008
16. Gray K, Graham J (2018) Atlas of moral psychology. The Guilford Press, New York
17. Voiklis J, Malle B (2017) Moral cognition and its basis in social cognition and social regulation. In: Gray K, Graham J (eds) Atlas of moral psychology. The Guilford Press, New York, pp 108–120
18. Christensen JF, Gomila A (2012) Moral dilemmas in cognitive neuroscience of moral decision-making: a principled review. Neurosci Biobehav Rev 36(4):1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008
19. Greene JD, Haidt J (2002) How (and where) does moral judgment work? Trends Cogn Sci 6:517–523
20. Laakasuo M, Sundvall J (2016) Are utilitarian/deontological preferences unidimensional? Front Psychol 7:1228. https://doi.org/10.3389/fpsyg.2016.01228
21. Paxton JM, Ungar L, Greene JD (2012) Reflection and reasoning in moral judgment. Cogn Sci 36:163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x
22. Miller G (2008) The roots of morality. Science 320:734–738
23. Wallach W, Allen C (2009) Moral machiens teaching robots right from wrong. Oxford University Press, Oxford
24. Cushman F, Greene JD (2012) Finding faults: how moral dilemmas illuminate cognitive structure. Soc Neurosci 7(3):269–279
25. Wolkenstein A (2018) What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. Ethics Inf Technol 20:163–173. https://doi.org/10.1007/s10676-018-9456-6
26. Kurzban R, DeScioli P, Fein D (2012) Hamilton versus Kant: pitting adaptations for altruism against adaptations for moral judgment. Evol Hum Behav 33:323–333. https://doi.org/10.1016/j.evolhumbehav.2011.11.002
27. Uhlmann EL, Pizarro DA, Diermeier D (2015) A person-centered approach to moral judgment. Perspect Psychol Sci 10:72–81. https://doi.org/10.1177/1745691614556679
28. Swann WB, Gómez Á, Dovidio JF, Hart S, Jetten J (2010) Dying and killing for one's group: identity fusion moderates responses to intergroup versions of the trolley problem. Psychol Sci 21:1176–1183. https://doi.org/10.1177/0956797610376656
29. Baron J, Jost JT (2019) False equivalence: are liberals and conservatives in the USA equally biased? Perspect Psychol Sci 14:292–303. https://doi.org/10.1177/1745691618788876
30. Blau A (2009) Hobbes on corruption. Hist Polit Thought 30(4):596–616
31. Köbis NC, van Prooijen J-W, Righetti F, Van Lange PAM (2016) Prospection in individual and interpersonal corruption dilemmas. Rev Gen Psychol 20:71–85. https://doi.org/10.1037/gpr0000069
32. Ostrom E (2000) Collective action and the evolution of social norms. J Econ Perspect 14:137–158. https://doi.org/10.1257/jep.14.3.137
33. Jost JT, Kay AC (2010) Social justice. In: Handbook of social psychology. Wiley, pp 1122–1165
34. Duke AA, Begue L (2015) The drunk utilitarian: blood alcohol concentration predicts utilitarian responses in moral dilemmas. Cognition 134:121–127. https://doi.org/10.1016/j.cognition.2014.09.006
35. Khemiri L, Guterstam J, Franck J, Jayaram-Lindström N (2012) Alcohol dependence associated with increased utilitarian moral judgment: a case control study. PLoS ONE 7:e39882. https://doi.org/10.1371/journal.pone.0039882
36. Descioli P, Kurzban R (2013) A solution to the mysteries of morality. Psychol Bull 139:477–496. https://doi.org/10.1037/a0029065
37. DeScioli P, Kurzban R (2009) Mysteries of morality. Cognition 112:281–299. https://doi.org/10.1016/j.cognition.2009.05.008
38. Cikara M, Farnsworth RA, Harris LT, Fiske ST (2010) On the wrong side of the trolley track: neural correlates of relative social valuation. Soc Cogn Affect Neurosci 5:404–413. https://doi.org/10.1093/scan/nsq011
39. Gray K, Young L, Waytz A (2012) Mind perception is the essence of morality. Psychol Inq 23:101–124. https://doi.org/10.1080/1047840X.2012.651387
40. Malle BF, 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. ACM, pp 117–124. https://doi.org/10.1145/2696454.2696458
41. Bigman YE, Waytz A, Alterovitz R, Gray K (2019) Holding robots responsible: the elements of machine morality. Trends Cogn Sci 23(5):365–368. https://doi.org/10.1016/j.tics.2019.02.008
42. Breazeal C, Gray J, Hoffman G, Berlin M (2004) Social robots: beyond tools to partners. In: RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759), pp 551–556
43. Coeckelbergh M (2011) Humans, animals, and robots: a phenomenological approach to human-robot relations. Int J Soc Robot 3:197–204. https://doi.org/10.1007/s12369-010-0075-6
44. Kulić CD, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):7–81. https://doi.org/10.1007/s12369-008-0001-3
45. Kumar EV, Li X, Sollers J, Stafford RQ, MacDonald BA, Wegner DM (2013) Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. PLoS ONE 8:e72589. https://doi.org/10.1371/journal.pone.0072589
46. Chee BTT, Taezoon P, Xu Q, Ng J, Tan O (2012) Personality of social robots perceived through the appearance. Work 41:272–276. https://doi.org/10.3233/WOR-2012-0168-272
47. Dubal S, Foucher A, Jouvent R, Nadel J (2011) Human brain spots emotion in non humanoid robots. Soc Cogn Affect Neurosci 6:90–97
48. Melson GF, Kahn PH, Beck A, Friedman B, Roberts T, Garrett E, Gill BT (2009) Children's behavior toward and understanding of robotic and living dogs. J Appl Dev Psychol 30:92–102. https://doi.org/10.1016/j.appdev.2008.10.011
49. Mori M (1970) The uncanny valley. Energy 7:33–35
50. MacDorman KF, Green RD, Ho CC, Koch CT (2009) Too real for comfort? uncanny responses to computer generated faces. Comput Human Behav 25:695–710. https://doi.org/10.1016/j.chb.2008.12.026

51. Mitchell WJ, Szerszen KA, Lu AS, Schermerhorn PW, Scheutz M, MacDorman KF (2011) A mismatch in the human realism of face and voice produces an uncanny valley. i-Perception 2(1):10–12. https://doi.org/10.1068/i0415
52. MacDorman KF (2005) Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it. In CogSci-2005 Workshop: toward social mechanisms of android science
53. Ferrey AE, Burleigh TJ, Fenske MJ (2015) Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. Front Psychol 6:249. https://doi.org/10.3389/fpsyg.2015.00249
54. Palomäki J, Kunnari A, Drosinou M, Koverola M, Lehtonen N, Halonen J, Repo M, Laakasuo M (2018) Evaluating the replicability of the uncanny valley effect. Heliyon 4:e00939. https://doi.org/10.1016/j.heliyon.2018.e00939
55. Sanfey A, Rilling J, Aronson J (2003) Neural basis of economic decision-making in the ultimatum game. Science 300:1755–1758
56. Aharoni E, Fridlund AJ (2007) Social reactions toward people vs. computers: how mere lables shape interactions. Comput Human Behav 23:2175–2189. https://doi.org/10.1016/j.chb.2006.02.019
57. Guadagno RE, Blascovich J, Bailenson JN, Mccall C (2007) Virtual humans and persuasion: the effects of agency and behavioral realism. Media Psychol 10:1–22.
58. Hoyt CL, Blascovich J, Swinth KR (2003) Social inhibition in immersive virtual environments. Presence Teleoperators Virtual Environ 12(2):183–195. https://doi.org/10.1162/1054746033 21640932
59. Nowak KL, Biocca F (2003) The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. In Presence: teleoperators and virtual environments, pp 481–494. https://doi.org/10.1162/105474603322761289
60. Kahn PH, Kanda T, Ishiguro H, Freier NG, Severson RL, Gill BT, Ruckert JH, Shen S (2012) "Robovie, you'll have to go into the closet now": children's social and moral relationships with a humanoid robot. Dev Psychol 48:303–314. https://doi.org/10.1037/a0027033
61. Severson RL, Carlson SM (2010) Behaving as or behaving as if? Children' s conceptions of personified robots and the emergence of a new ontological category. Neural Netw 23:1099–1103. https://doi.org/10.1016/j.neunet.2010.08.014
62. Chalmers DJ (2010) The singularity: a philosophical analysis. J Conscious Stud 17:7–65. https://doi.org/10.1002/9781118922590.ch16
63. Boyer P, Barrett C (2005) Evolved intuitive ontology: integrating neural, behavioral and developmental aspects of domain-specificity. In: Buss DM (ed) Handbook of evolutionary psychology. Wiley, New York
64. Putt SS, Wijeakumar S, Franciscus RG, Spencer JP (2017) The functional brain networks that underlie Early Stone Age tool manufacture. Nat Hum Behav 1(6):1–8. https://doi.org/10.1038/s41562-017-0102
65. Atran S, Medin D, Ross N (2004) Evolution and devolution of knowledge: a tale of two biologies. J R Anthropol Inst 10:395–420. https://doi.org/10.1111/j.1467-9655.2004.00195.x
66. Amiot CE, Bastian B (2015) Toward a psychology of human-animal relations. Psychol Bull 141:6–47. https://doi.org/10.1037/a0038147
67. Billings DR, Schaefer KE, Chen JYC, Hancock PA (2012) Human-robot interaction: developing trust in robots. In HRI'12-proceedings of the 7th annual ACM/IEEE international conference on human-robot interaction, pp 109–110. https://doi.org/10.1145/2157689.2157709
68. Borenstein J, Howard A, Wagner A (2017) Pediatric robotics and ethics: the robot is ready to see you now, but should it be trusted? In: Lin P, Jenkins R, Abney K (eds) Robot ethics 2.0. Oxford University Press, Oxford, pp 127–141
69. Moore RK (2012) A Bayesian explanation of the "Uncanny Valley" effect and related psychological phenomena. Sci Rep 2:864. https://doi.org/10.1038/srep00864
70. Saygin AP, Chaminade T, Ishiguro H, Driver J, Frith C (2012) The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Soc Cogn Affect Neurosci 7:413–422. https://doi.org/10.1093/scan/nsr025
71. Gawronski B, Beer JS (2016) What makes moral dilemma judgments "utilitarian" or "deontological"? Soc Neurosci 12(6):626–632. https://doi.org/10.1080/17470919.2016.1248787
72. Kunnari A, Sundvall J, Laakasuo M (2020) Challenges in process dissociation measures for moral cognition. Front Psychol 11:559934. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.559934
73. Ward AF, Olsen AS, Wegner DM (2013) The harm-made mind: observing victimization augments attribution of minds to vegetative patients, robots, and the dead. Psychol Sci 24(8):1437–1445
74. Wilson Van Voorhis CR, Morgan BL (2007) Understanding power and rules of thumb for determining sample sizes. Tutor Quant Methods Psychol 3:43–50
75. Greene JD (2018) Can we understand moral thinking without understanding thinking? In Atlas of moral psychology, pp 3–8
76. Moravec H (1988) Mind children: the future of robot and human intelligence. Harvard University Press, Harvard
77. Tooby J, Cosmides L (1992) The psychological foundations of culture. In The adapted mind: evolutionary psychology and the generation of culture, pp 19–49
78. Laakasuo M, Sundvall J, Berg A, Drosinou M-A, Herzon V, Kunnari AJO, Koverola M, Repo M, Saikkonen T, Palomäki JP (2021) Moral psychology and artificial agents (part 1): ontologically categorizing bio-cultural humans. In: Machine law, ethics, and morality in the age of artificial intelligence. IGI Global, pp 166–188
79. Laakasuo M, Sundvall J, Berg A, Drosinou M-A, Herzon V, Kunnari AJO, Koverola M, Repo M, Saikkonen T, Palomäki JP (2021) Moral psychology and artificial agents (part 2): The transhuman connection. In: Machine law, ethics, and morality in the age of artificial intelligence. IGI Global, pp 189–204

Michael Laakasuo is an adjunct professor of cognitive sciences at University of Helsinki. He iscurrently the PI of Moralities of Intelligent Machines (www.moim.fi) -research project funded byJane and Aatos Erkko Foundation and Academy of Finland.

Jussi Palomäki is an adjunct professor of cognitive science at University of Helsinki and a part time lecturer. He has studied human-computer interaction and decision-making in gaming environments. He is currently working on issues related to related to high performance cognition and betting behavior.

Nils Köbis is a Post-Doc at the Max-Planck-Institute for Human Development (Center for Humans & Machine), where he is conducting research on "(Un)ethical Behavior of Humans and Machines".