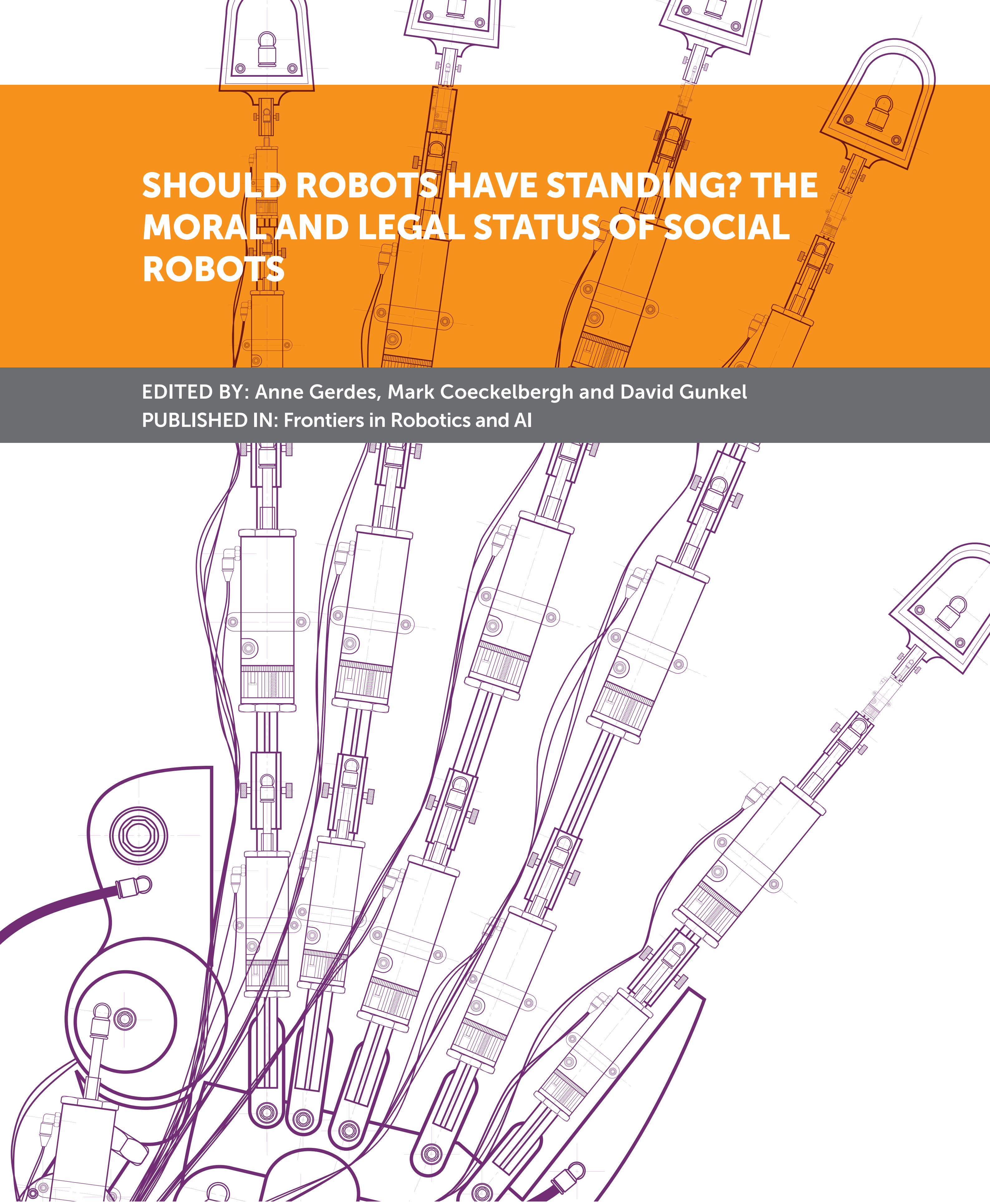

SHOULD ROBOTS HAVE STANDING? THE MORAL AND LEGAL STATUS OF SOCIAL ROBOTS

EDITED BY: Anne Gerdes, Mark Coeckelbergh and David Gunkel

PUBLISHED IN: Frontiers in Robotics and AI

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-537-9
DOI 10.3389/978-2-88976-537-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

SHOULD ROBOTS HAVE STANDING? THE MORAL AND LEGAL STATUS OF SOCIAL ROBOTS

Topic Editors:
Anne Gerdes, University of Southern Denmark, Denmark
Mark Coeckelbergh, University of Vienna, Austria
David Gunkel, Northern Illinois University, United States

Citation: Gerdes, A., Coeckelbergh, M., Gunkel, D., eds. (2022). Should Robots Have Standing? The Moral and Legal Status of Social Robots. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-537-9

Table of Contents

frontiers | Frontiers in Robotics and AI

EDITORIAL
published: 16 June 2022
doi: 10.3389/frobt.2022.946529

Editorial: Should Robots Have Standing? The Moral and Legal Status of Social Robots

David J. Gunkel[1]*, *Anne Gerdes*[2] *and Mark Coeckelbergh*[3]

[1]*Department of Communication, Northern Illinois University, DeKalb, IL, United States,* [2]*Department of Design and Communication, University of Southern Denmark, Odense, Denmark,* [3]*Department of Philosophy, University of Vienna, Vienna, Austria*

Keywords: social robot, artificial intelligence, ethics, law, philosophy of technology, AI ethics, AI and society, technology policy

Editorial on the Research Topic

Should Robots Have Standing? The Moral and Legal Status of Social Robots

In a proposal issued by the European Parliament (Delvaux, 2016) it was suggested that robots might need to be considered "electronic persons" for the purposes of social and legal integration. The very idea sparked controversy, and it has been met with both enthusiasm and resistance. Underlying this disagreement, however, is an important moral/legal question: When (if ever) would it be necessary for robots, AI, or other socially interactive, autonomous systems to be provided with some level of moral and/or legal standing?

This question is important and timely because it asks about the way that robots will be incorporated into existing social organizations and systems. Typically technological objects, no matter how simple or sophisticated, are considered to be tools or instruments of human decision making and action. This instrumentalist definition (Heidegger, 1977; Feenberg, 1991; Johnson, 2006) not only has the weight of tradition behind it, but it has so far proved to be a useful method for responding to and making sense of innovation in artificial intelligence and robotics. Social robots, however, appear to confront this standard operating procedure with new and unanticipated opportunities and challenges. Following the predictions developed in the computer as social actor studies and the media equation (Reeves and Nass, 1996), users respond to these technological objects as if they were another socially situated entity. Social robots, therefore, appear to be more than just tools, occupying positions where we respond to them as another socially significant Other.

This Research Topic of *Frontiers in Robotics* seeks to make sense of the social significance and consequences of technologies that have been deliberately designed and deployed for social presence and interaction. The question that frames the issue is "Should robots have standing?" This question is derived from an agenda-setting publication in environmental law and ethics written by Christopher Stone, *Should Trees Have Standing? Toward Legal Rights for Natural Objects* (1974). In extending this mode of inquiry to social robots, contributions to this Research Topic of the journal will 1) debate whether and to what extent robots can or should have moral status and/or legal standing, 2) evaluate the benefits and the costs of recognizing social status, when it involves technological objects and artifacts, and 3) respond to and provide guidance for developing an intelligent and informed plan for the responsible integration of social robots.

In order to address these matters, we have assembled a team of fifteen researchers from across the globe and from different disciplines, who bring to this conversation a wide range of viewpoints and methods of investigation. These contributions can be grouped and organized under the following four subject areas:

OPEN ACCESS

Edited and reviewed by:
Bertram F. Malle,
Brown University, United States

****Correspondence:***
David J. Gunkel
dgunkel@niu.edu

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

Received: *17 May 2022*
Accepted: *30 May 2022*
Published: *16 June 2022*

Citation:
Gunkel DJ, Gerdes A and Coeckelbergh M (2022) Editorial: Should Robots Have Standing? The Moral and Legal Status of Social Robots.
Front. Robot. AI 9:946529.
doi: 10.3389/frobt.2022.946529

STANDING AND LEGAL PERSONALITY

Five of the essays seek to take-up and directly address the question that serves as the title to this special issue: Should robots have standing? In "Speculating About Robot Moral Standing: On the Constitution of Social Robots as Objects of Governance" Jesse De Pagter argues that the question of robot standing—even if it currently is a future-oriented concern and speculative idea—is an important point of discussion and debate in the critical study of technology. His essay therefore situates social robot in the context of anticipatory technology governance and explains how a concept like robot standing informs and can be of crucial importance to the success of this endeavor.

"Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence," Brazilian jurist Avila Negri performs a cost/benefit analysis of legal proposals like that introduced by the European Parliament. In his reading of the existing documents, Avila Negri finds evidence of a legal pragmatism that seeks guidance from the precedent of corporate law but unfortunately does so without taking into account potential problems regarding the embodiment of companies and the specific function of the term "legal person" in the grammar of law.

In "Robots and AI as Legal Subjects? Disentangling the Ontological and Functional Perspective," Bertolini and Episcopo seek to frame and formulated a more constructive method for deciding the feasibility of granting legal standing to robotic systems. Toward this end, they argue that *standing* should be strictly understood as a legal affordance such that the attribution of subjectivity to an artifact needs to be kept entirely within the domain of law, and grounded on a functional, bottom-up analysis of specific applications. Such an approach, they argue, usefully limits decisions about moral and legal status to practical concerns and legal exigencies instead of getting mired in the philosophical problems of attributing animacy or agency to artifacts.

These two efforts try to negotiate the line that distinguishes what is a thing from who is a person. Other contributions seek to challenge this mutually exclusive dichotomy by developing alternatives. In "The Virtuous Servant Owner—A Paradigm Whose Time has Come (Again)," Navon introduces a third category of entity, a kind of in between status that is already available to us in the ancient laws of slavery. Unlike other proposals that draw on Roman law, Navon formulates his alternative by turning to the writings of the Jewish philosopher Maimonides, and he focuses attention not on the legal status of the robot-slave but on the moral and legal opportunities imposed on its human master.

In "Gradient Legal Personhood for AI Systems—Painting Continental Legal Shapes Made to Fit Analytical Molds" Mocanu proposes another solution to the person/thing dichotomy that does not—at least not in name—reuse ancient laws of slavery. Instead of trying to cram robots and AI into one or the other of the mutually exclusive categories of person or thing, Mocanu proposes a gradient theory of personhood, which employs a more fine-grained spectrum of legal statuses that does not require one to make simple and limited either/or distinctions between legal subjects and objectivized things.

PUBLIC OPINION AND PERCEPTION

Deciding these matters is not something that is or even should be limited to legal scholars and moral philosophers. These are real questions that are beginning to resonate for users and non-experts. The contribution from the Dutch research team of Graaf et al. explores a seemingly simple and direct question: "Who Wants to Grant Robots Rights?" In response to this question, they survey the opinions of non-expert users concerning a set of specific rights claims that have been derived from existing international human rights documents. In the course of their survey, they find that attitudes toward granting rights to robots largely depend on the cognitive and affective capacities people *believe* robots possess or will possess in the future.

In "Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection," Martínez and Winter investigate a similar question: To what extent, if any, should the law protect sentient artificial intelligence? Their study, which was conducted with adults in the United States, found that only one third of survey participants are likely to endorse granting personhood and standing to sentient AI (assuming its existence), meaning that the majority of the human subjects they surveyed are not—at least not at this point in time—in favor of granting legal protections to intelligent artifacts. These finding are consistent with an earlier study that the authors conducted in 2021 with legal professionals.

SUFFERING AND MORAL/LEGAL STATUS

Animal rights philosophy and many animal welfare laws derive from an important conceptual innovation attributed to the English political philosopher Jeremy Bentham. For Bentham what mattered and made the difference for moral and legal standing was not the usual set of human-grade capacities, like self-consciousness, rationality, or language use. It was simply a matter of sentience: "The question is not, 'Can they reason?' nor, 'Can they talk?' but 'Can they suffer?'" (Bentham 2005, 283). For this reason, the standard benchmark for deciding questions of moral and legal standing—a way of dividing who is a person from what remains a mere thing—is an entity's ability to suffer or to experience pain and pleasure. And several essays leverage this method in constructing their response to the question "should robots have standing?"

In the essay "From Warranty Voids to Uprising Advocacy: Human Action and the Perceived Moral Patiency of Social Robots," Banks employs a social scientific study to investigate human users' perceptions of the moral status of social robots. And she finds significant evidence that people can imagine clear dynamics by which robots may be said to benefit and suffer at the hands of humans.

In "Whether to Save a Robot or a Human: On the Ethical and Legal Limits of Protections for Robots," legal scholar Mamak investigates how this human-all-too-human proclivity for concern with robot well-being and suffering might run afoul of the law, which typically prioritizes the welfare of human subjects and even stipulates the active protection of humans over other kind of things. In effect, Mamak critically evaluates the legal contexts and consequences of the social phenomena that has been reported in empirical studies like that conducted by Banks.

And with the third essay in this subject area, "The Conflict Between People's Urge to Punish AI and Legal Systems," Lima et al. explore the feasibility of extending legal personhood to AI and robots by surveying human beings' perceptions of liability and punishment. Data from their inventory identifies a conflict between the desire to punish automated agents for wrongful action and the perceived impracticability of doing so when the agent is a robot or AI lacking conscious experience.

RELATIONAL ETHICS

In both moral philosophy and law, what something is largely determines its moral and legal status. This way of proceeding, which makes determinations of standing dependent on ontological or psychological properties, like consciousness or sentience, has traction in both moral philosophy and law. But it is not the only, or even the best, method for deciding these matters. One recent and promising alternative is relational ethics. The final set of essays investigate the opportunities and challenges of this moral and legal innovation.

In "Empathizing and Sympathizing With Robots: Implications for Moral Standing" Quick employs a phenomenological approach to investigating human-robot interaction (HRI), arguing that empathetic and sympathetic engagements with social robots takes place in terms of and is experienced as an ethical encounter. Following from this, Quick concludes, such artifacts will need to be recognized as another form of socially significant otherness and would therefore be due a minimal level of moral consideration.

With "Robot Responsibility and Moral Community," Dane Leigh Gogoshin recognizes that the usual way of deciding questions of moral responsibility would certainly exclude robots due to the fact that these technological artifacts lack the standard qualifying properties to be considered legitimate moral subjects, i.e. consciousness, intentionality, empathy, etc. But, Gogoshin argues, this conclusion is complicated by actual moral responsibility practices, where human beings often respond to rule-abiding robots as morally responsible subjects and thus members of the moral community. To address this, Gogoshin proposes alternative accountability structures that can accommodate these other forms of moral agency.

The essay "Does the Correspondence Bias Apply to Social Robots?: Dispositional and Situational Attributions of Human Versus Robot Behavior" adds empirical evidence to this insight. In this essay, human-machine communication researchers Edwards and Edwards investigate whether correspondence bias (e.g. the tendency for individuals to over-emphasize personality-based explanations for other people's behavior while under-emphasizing situational explanations) applies to social robots. Results from their experimental study indicate that participants do in fact make correspondent inferences when evaluating robots and attribute behaviors of the robot to perceived underlying attitudes even when such behaviors are coerced.

With the essay "On the Social-Relational Moral Standing of AI: An Empirical Study Using AI-Generated Art," Lima et al. turn attention from the social circumstances of HRI to a specific domain where robot intervention is currently disrupting expected norms. In their social scientific investigation, the authors test whether and how interacting with AI-generated art affects the perceived moral standing of its creator, and their findings provide useful and empirically grounded insights concerning the operative limits of moral status attribution.

Finally, if these three essays provide support for a socially situated form of relational ethics, then the essay from Sætra—"Challenging the Neo-Anthropocentric Relational Approach to Robot Rights"—provides an important counterpoint. Unlike traditional forms of moral thinking where what something is determines how it is treated, relationalism promotes an alternative procedure that flips the script on this entire transaction. In his engagement with the existing literature on the subject, Sætra finds that the various articulations of "relationalism," despite many advantages and opportunities, might not be able to successfully resolve or escape from the problems that have been identified.

In presenting this diverse set of essays, our intention has been to facilitate and stage a debate about the moral and legal status of social robots that can help theorists and practitioners not only make sense of the current state of research in this domain but also assist them in the development of their own thinking about and research into these important and timely concerns. Consequently, our objective with the Research Topic is not to advance one, definitive solution or promote one way to resolve these dilemmas but to map the range of possible approaches to answering these questions and provide the opportunity for readers to critically evaluate their significance and importance.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Bentham, J. (2005). *An Introduction to the Principles of Morals and Legislation*. New York: Oxford University Press.

Delvaux, M. (2016). *Draft Report, with Recommendations to the Commission on Civil Law Rules on Robotics, 2015/2103(INL). Committee on Legal Affairs.* Brussel: European Parliament. https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect.

Feenberg, A. (1991). *Critical Theory of Technology*. New York: Oxford University Press.

Heidegger, M. (1977). "The Question Concerning Technology and Other Essays," in *Trans. W. Lovitt* (New York: Harper & Row). Originally published 1962.

Johnson, D. G. (2006). Computer Systems: Moral Entities but Not Moral Agents. *Ethics Inf. Technol.* 8, 195–204. doi:10.1007/s10676-006-9111-5

Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge: Cambridge University Press.

Stone, C. D. (1974). *Should Trees Have Standing? toward Legal Rights for Natural Objects*. Los Altos, CA: William Kaufmann.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 28 May 2021
doi: 10.3389/frobt.2021.670503

From Warranty Voids to Uprising Advocacy: Human Action and the Perceived Moral Patiency of Social Robots

*Jaime Banks**

College of Media & Communication, Texas Tech University, Lubbock, TX, United States

Moral status can be understood along two dimensions: moral agency [capacities to be and do good (or bad)] and moral patiency (extents to which entities are objects of moral concern), where the latter especially has implications for how humans accept or reject machine agents into human social spheres. As there is currently limited understanding of how people innately understand and imagine the moral patiency of social robots, this study inductively explores key themes in how robots may be subject to humans' (im)moral action across 12 valenced foundations in the moral matrix: care/harm, fairness/unfairness, loyalty/betrayal, authority/subversion, purity/degradation, liberty/oppression. Findings indicate that people can imagine clear dynamics by which anthropomorphic, zoomorphic, and mechanomorphic robots may benefit and suffer at the hands of humans (e.g., affirmations of personhood, compromising bodily integrity, veneration as gods, corruption by physical or information interventions). Patterns across the matrix are interpreted to suggest that moral patiency may be a function of whether people diminish or uphold the ontological boundary between humans and machines, though even moral upholdings bare notes of utilitarianism.

Keywords: Moral patiency, mental models, ontological categorization, morphology, boundary objects

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Dennis Küster,
University of Bremen, Germany
Chad Edwards,
Western Michigan University,
United States

****Correspondence:***
Jaime Banks
j.banks@ttu.edu

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *21 February 2021*
Accepted: *12 May 2021*
Published: *28 May 2021*

Citation:
Banks J (2021) From Warranty Voids to Uprising Advocacy: Human Action and the Perceived Moral Patiency of Social Robots. Front. Robot. AI 8:670503. doi: 10.3389/frobt.2021.670503

INTRODUCTION

The Tin Woodman of Oz fame (alias Nick Chopper) was an autonomous, metal-made man—a robot of sorts. In one account (Baum, 1904), Chopper was fast but far-flung friends with Scarecrow. Scarecrow traveled with companions to see Chopper, and warned those companions to refer to Chopper as Emperor to honor his authority. Upon arrival, Scarecrow offers Chopper a warm greeting and embrace as a matter of care. Chopper realizes that he is in poor condition for company and seeks servants to polish him to a pure sheen. Finally, Scarecrow gives Chopper fair warning of incoming invaders bound to threaten his domain's freedom. In this way, Chopper—a machine agent—was afforded moral considerations of care, fairness, authority, loyalty, purity, and liberty.

Contemporary machine agents may also be the target of human moral consideration, in both positive forms (e.g., accommodating Roomba robots; Sung et al., 2007) and negative (e.g., physical abuse of hitchBOT; Grodzinsky et al., 2019). However, empirical inquiries into moral status of social machines tend to focus narrowly on notions of morality when attention to the full moral matrix is warranted—inclusive of care/harm, fairness/unfairness, authority/subversion, loyalty/betrayal, purity/degradation, and liberty/oppression, as laid out by Moral Foundations Theory (Graham et al., 2011; Iyer et al., 2012). Thus, there is a knowledge gap often filled by a tendency to rely on human moral norms to consider machine moral dynamics. This investigation aims to begin addressing that gap by identifying understandings of how social robots may be considered patients to humans' (im)moral actions. In other words: In what ways do people innately see social robots as (un)deserving of moral consideration, and how do people imagine those dynamics playing out in everyday life? Answering this question is necessary as we must have a holistic, empirically grounded grasp on the nature of machines' perceived moral status before we may meaningfully understand its implications—working to know what it *is* before we can fully understand why and how it *matters*. To this end, I conducted an inductive thematic analysis of elicited stories regarding social robots' moral patiency to human action. Findings indicate that people see rich and varied potentials for machine moral patiency across the moral matrix; robots' moral patiency appears to rest largely on how humans recognize or reject their personhood by upholding or diminishing the human/machine ontological boundary.

LITERATURE REVIEW

Agents' moral status may be understood as having two primary dimensions: moral agency and moral patiency. Moral agency is the capacity to be good and do good (Banks, 2020a) and its relevance to robots has received ample attention in extant literature. Less attention has been paid to moral patiency—the ways in which robots may be victims or beneficiaries of (im)moral action (Gunkel, 2018).

Social Robots as Moral Patients

Moral events requires both agents (intentional actors) and patients (targets of action; Gray and Wegner, 2009). A moral patient is an entity that can and/or should be the object of moral concern such that others must account for its interests (Anderson, 2013). Whereas moral agents manifest autonomy and intentionality, moral patients cannot necessarily decide to act such that the actor (or society, broadly) is responsible for preserving the patient's well-being (see Bryson, 2018). Moral patiency, then, is a state of holding unintentional-subject status to some degree.

The qualifiers of *can* and *should* are key in considering whether an entity—here, a social robot—may be assigned moral-patient status (Gunkel, 2018). Whether a robot can be a patient is an operational question: Does it have capabilities or properties that create the conditions for moral patiency? Often, qualifying characteristics are anthropogenic properties such as self-interestedness, emotion, or consciousness (Coeckelbergh, 2010). Sometimes they are more general properties like autonomy, interactivity (Coeckelbergh, 2020), or goal-directedness (Anderson, 2013). More generally, moral patiency is thought to require the capacity to feel pain or pleasure (Sparrow, 2004). In turn, whether a robot should be a moral patient is an ethical question: Are they due moral consideration by virtue of some (in)direct obligation? In this question, direct consideration is warranted when the target has some inherent value, and indirect consideration is based on some extrinsic value (Coeckelbergh, 2020; Friedman, 2020). For some, a robot need not meet anthropocentric criteria, as they may have some phenomenal processes analogous to emotion or self-awareness, where the processes are qualifying but humans are unable to detect them (Davenport, 2014; Coeckelbergh, 2020).

Notably, some argue that it is inherently immoral to ascribe moral status to a robot given that robots would then have to compete with humans for resources or other forms of status (Bryson, 2018). Others still suggest these are moot points because humans will ultimately not be the arbiters of robots' moral standing as AI advances and robots may eventually demand their rights, as have other subjugated groups (Asaro, 2006). It is beyond the scope of this work to take a position on the criteria for questions of can or should. Rather, it is to focus on human *perception* of robot's potential moral patiency.

Robots as Perceived Moral Patients

As noted, being a moral patient is an operational status while deserving to be one is an ethical valuation. There is, however, a third facet of moral status that requires attention: the degree to which an entity is *perceived* to be an object of moral concern, irrespective of whether it operationally can be or ethically ought to be. In my past work (e.g., Banks et al., 2021), experimentally manipulating robot behaviors to induce certain reactions has proven unreliable, with perceptions of the behaviors proving more powerful than the form of the behaviors. This is likely due to variation in people's understandings of what robots are and how they work (Banks, 2020b), since those understandings (i.e., mental models; Craik, 1943) shape actual or imagined experiences. So while the perceived moral patient (PMP)[1] can be conceptualized as an entity that is thought to be an object of moral concern, it may be operationalized as an entity for which an observer's mental model contains some belief that the entity can benefit or suffer at the hands of others. The moral-patient status of a robot, from this frame, is not an adjudication on the nature of the robot itself (whether it can or should be considered) but instead on the subjective orientation of a human as they imagine or observe the robot existing among humans (cf. Coeckelbergh, 2018). The robot-as-PMP effectively exists in the human mind, manifested in mental models for robots, and this cluster of ideas guides the ways that human may consider (a/im)morally engaging robots in actual encounters.

The robot-as-PMP could be said to emerge through observations or inferences of a robot's particular properties (e.g., sentience, free will; Coeckelbergh, 2012), mental status (Gray et al., 2012), benefit or suffering experiences (Sparrow, 2004), or personal histories (Darling et al., 2015). However, these are all attendant to the robot, while PMPs reside in the subjective experience of the observing or imagining human. Thus, it is prudent to explore the nature of the robot-as-PMP by examining people's held ideas about the moral relations between humans

[1] I use the abbreviation PMP throughout to refer both to perceived moral patiency (the state of an entity, having had the particular moral standing ascribed to it) and the perceived moral patient (the entity itself, as it manifests in the mind of the perceiver), with the specific meaning indicated by context.

and machines, in line with a social-relational approach that "takes seriously the phenomenology and experience of other entities such as robots… (such that) the robot may appear as a quasi-other; this turns the question about 'status' into a question about social relations…" (Coeckelbergh, 2018, p. 149).

Understanding the robot-as-PMP can be challenging in that mental models are proverbial black boxes (Rouse and Morris, 1985) and people often will not overtly ascribe moral status to robots as they would to humans despite judging their behaviors as similarly good or bad (Banks, 2020a). It is useful, then, to draw on moral typecasting theory (MTT; Gray and Wegner, 2009) in tandem with eliciting hypothetical stories to infer mental-model content (cf. de Graaf and Malle, 2019). MTT contends that perceptions of moral agency and patiency are inversely related such that—in dyadic moral relations—as one entity is seen as more of an agent, the other is seen as more of a patient; this dynamic manifests across naturally varying degrees of agency/patiency, across moral valence (good/bad), and in both causal directions (perceived agency influences perceived patiency and vice-versa; Gray and Wegner, 2009). Through this lens, agency and patiency are not categories of actors but instead matters of asymmetrical degree. Thus, if robots' PMP status is inversely related to the humans' perceived moral-agent status, the robot-as-PMP may be understood by identifying patterns in humans' ideas about human action toward robots.

Situating Robot-as-PMP Within the Moral Matrix

Perceptions of robots' PMP status have been empirically examined, but often in a narrow fashion and often with assumptions that human norms are neatly applied to the moral standing of robots. In contrast, Moral Foundations Theory (MFT) argues that moral judgments are intuited across a matrix of modules (i.e., foundations): care/harm, fairness/unfairness, loyalty/betrayal, purity/degradation, authority/subversion (Graham et al., 2011), and the candidate foundation liberty/oppression (Iyer et al., 2012). In considering current understandings of robots-as-PMPs, it would be beyond the scope of this project to offer a comprehensive review, however it is prudent to offer a brief encapsulation of empirical works to highlight known social-psychological operations for each foundation. The following foundation definitions and their respective virtues are drawn from MFT's foundational works (Graham et al., 2011; Iyer et al., 2012)

Care/Harm

The care foundation (violation: harm) accounts for the physical or psychological pains and pleasures experienced by others, with liking of pleasure and disliking of pain moving people to kindness and compassion. Examinations of care for robots are most evident in relation to empathy. People express greater empathy for highly anthropomorphic robots than for those with machinic morphologies (Riek et al., 2009) especially for physical-pain empathy (Chang et al., 2021). However, evidence of preconscious processing suggests that people may react to emotional expressions even from non-humanoid robots (Dubal et al., 2011) and people may react more strongly to robots' dramatic suffering (e.g., potential death) than to everyday patiency situations (Nijssen et al., 2020). Regarding robots suffering harm, mind perception and consideration of painful suffering are entangled. People are more verbally aggressive toward a robot when there are lesser attributions of mind (Keijsers and Bartneck, 2018), but it may instead be the observation of suffering that moves people to infer mind (Ward et al., 2013). Conversely, people may be more hesitant to torture or kill robots when the machines present narrative histories, though this response may be predicated on high trait empathy (Darling et al., 2015). Situational factors may also influence harm-based PMP, as interventions are more likely when bystander robots express sadness at abuse (Connolly et al., 2020) or when patients fights back (Bartneck and Keijsers, 2020).

Fairness/Unfairness

The fairness foundation (violation: inequity or cheating) engages altruistic reciprocity, with giving fair chances linked to justice and trustworthiness. Scholarly attention to equity-fairness to robots is limited. Children have articulated that robots deserve fair treatment (though not necessarily liberty or rights; Kahn et al., 2012) and people are more likely to see robots as deserving of fairness when their behaviors are autonomous (*versus* remotely controlled; Gary, 2014). More often, studies of (un)fairness focus on cheating in joint activities. People are more likely to cheat (characterized as disregarding instructions) when the robot has a neutral personality *versus* a friendly or authoritarian one (Maggi et al., 2020). Other studies take up the fairness-like construct of reciprocity *via* ultimatum and prisoner's dilemma games. For instance, people may engage in more profitable, reciprocal collaborations when agents (including robots) engage in tit-for-tat strategies *versus* other approaches (Sandoval et al., 2016). However, such studies often characterize fair negotiation less as a moral question and more as a strategy or indicative of discrete psychological processes.

Loyalty/Betrayal

The loyalty foundation (violation: betrayal) encompasses the bonds inherent to coalitions (tribes, families, teams) that promote faithfulness, patriotism, and other group-affiliative virtues. This is often addressed as a matter of in-grouping/out-grouping based on teams or social-group signals. People prefer robots that signal similar cultural backgrounds (Trovato et al., 2015) or nationalities (Eyssel and Kuchenbrandt, 2012). Preference for ingroup robots over outgroup humans has been indicated by lower likelihood of inflicting discomfort to those robots (Fraune et al., 2017) and deference to a robot's instructions (Sembroski et al., 2017). Applied research accounts for how robot service providers (i.e., in hospitality and entertainment) may impact positively brand loyalty, however in those cases the robot is a mediator and the loyalty is to the brand rather than to the robot as a patient (e.g., Milman et al., 2020). An exception to this pattern takes up the telling of a robot's secrets as a violation of psychological intimacy (i.e., betrayal), finding that people were more likely to betray a robot's secret when the machine offered only rudimentary social cues *versus* more elaborately social cues (Kahn et al., 2015).

Authority/Subversion

The authority foundation (violation: subversion) includes deference to or undermining of institutional, functional, or principled superiors, as one may defer to others in acts of piety, obedience, or tradition. Deference to robots as authorities has been examined in several ways, though they are more a matter of functional trust and superior skill than as a matter of moral concern. For instance, the machine heuristic is a cognitive shortcut to the logic that: if machine, then systematic, accurate, and unbiased, therefore trustworthy (see Sundar, 2020). Operationalized as following instructions, people may be moved to disobey a robot when they feel its behavior is unsafe (Agrawal and Williams, 2017) or hesitate at (but ultimately obey) robots' directives that push moral boundaries (Aroyo et al., 2018). The anticipated effects of (dis)obedience may play a role as people will defer to humans over robots when instructions conflict and stakes are high (Sembroski et al., 2017).

Purity/Degradation

The purity foundation (violation: degradation) is an interesting module with respect to robots because it is definitionally tied to organic integrity that may not be seen as relevant to robots. Specifically, purity is characterized as aversion to contaminants or adulterations, where upholding purity manifests naturalness, chastity, or temperance virtues. Most closely related are forms of biological (im)purity, where robots are seen as subject to contamination (e.g., bacterial risks in healthcare contexts; Bradwell et al., 2020). They may also be made impure through use in antisocial sexual activities (e.g., satisfying rape fantasy; Cox-George and Bewley, 2018), although references to humans as the "actual victims" protected by therapeutic uses of robots suggests that robots may not be seen as meaningful patients (see Danaher, 2017). Notions of purity and degradation are discernible in discussions of metaphorical immune systems whereby robots may be kept pure by detecting non-self elements and diagnosing faults (Gong and Cai, 2008) such that degradation may emerge from viruses, breakage, or glitches.

Liberty/Oppression

Liberty (violation: oppression) is a candidate foundation encompassing rejection or engagement of controlling or dominating forces, where anti-control dispositions are associated with individualism and independence. Liberty may be linked to notions of rights, where having rights equates to an absence of oppression, where supporting robot rights is linked to prior attitudes toward machine agents (Spence et al., 2018). As the arguable default is for robots to be at the command of humans (i.e., oppressed thereby), notions of liberty/oppression are entangled with the moral questions around whatever a controlling human is asking a robot to do, such as forced sex with humans (degradation) or guarding of property (authority). However, liberty for robots may be seen as distinct from more general fair treatment (Kahn et al., 2012). United States populations generally disfavor assigning rights to robots; however, those attitudes may be based on misinformation about the legal nature of personhood (Lima et al., 2020).

Understanding Innate Perceptions of Robot Moral Patiency

The literature reviewed above (and broader coverage of human treatment of robots) is useful in understanding some of the moral mechanisms in human-machine interactions. However, attention to robot moral patiency generally suffers from several shortcomings. Empirical studies tend to a) rely on a priori judgments of what should matter in humans' considerations of robots without accounting for the mental models for morality and for robots that are brought into the encounters; b) rely on relatively narrow formulations of morality, often c) reflecting explicit, validated tests rather than messier worldly operations; d) application of human-patiency standards when they may not be relevant to robots; sometimes e) considering the patient conceptually or in isolation, removing the ostensible patient from the social context required for moral events to occur. These limitations result in a constrained understanding of how people see social machines as potential moral patients. To begin to address these constraints, it is necessary to (correspondingly) a) elicit imagined narratives of robots-as-PMPs b) across the moral matrix through c) native understandings of how moral events may play out, d) identifying foundation-specific conditions without constraining responses to human norms, e) positioned in the requisite social context of human-robot interaction whereby the robot may be patient to the human's agency. These requirements in mind, I ask (RQ1): How do people understand robot moral patiency as a function of human action?

METHOD

To address the research question, an online survey (N = 442) elicited descriptions of how humans may treat robots in moral and immoral ways. The study design relies on the notion that when people talk about the world in general and robots in particular, they relate narratives that externalize their internal understanding of the subject matter (de Graaf and Malle, 2019). Thus, elicited narratives may convey conceptions of robots as an "other" that may be engaged in moral relations (cf. Coeckelbergh, 2020), highlighting constructions of robots-as-PMPs. All study instrumentation, stimuli, data, and analysis-iteration narratives are available as online supplements at https://osf.io/5pdnc/.

Participants and Procedure

Participants comprised an approximately representative sample of United States residents (based on 2015 Census Bureau estimates for sex, race, age, and political ideology; see supplements for complete descriptives) empaneled through Prolific to participate in a 30-min online survey about "how robots might experience the world." Initial data were reviewed to ensure passing of attention checks, ensure clear address of the elicitation, and exclusion of nonsense and likely bot responses,

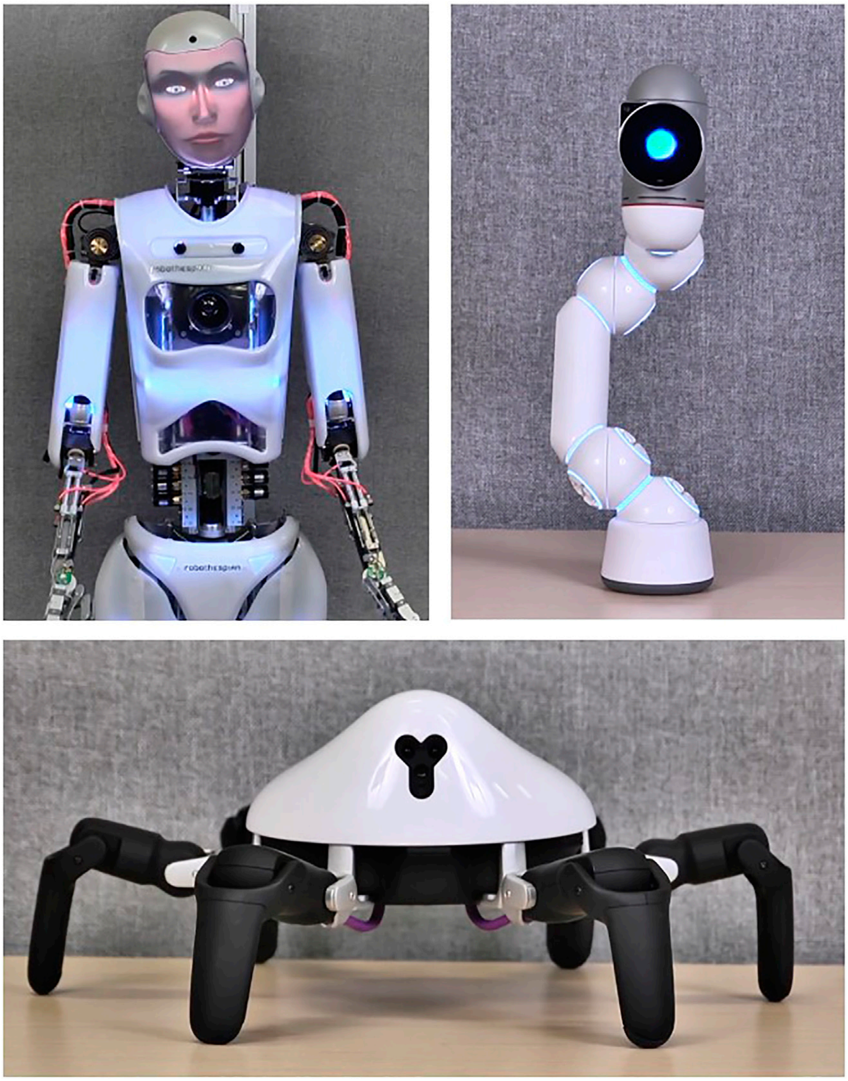

FIGURE 1 | Stimulus robots were anthropomorphic (Robothespian, top left), mechanomorphic (Clicbot, top right), and zoomorphic (Hexa, bottom).

resulting in $n = 43$ removals, and each was replaced according to sampling criteria.

After confirming informed consent, passing an audiovisual access check, and completing items capturing past experience with social robots, participants were randomly assigned to view a video of one of three robots, each offering an identical introduction. After giving first impressions of the robot, they were then randomly assigned to elicitations for three of the six MFT foundations inherent (limited to avoid fatigue); a third randomization then assigned an upholding or violating permutation for each of those three foundations. Importantly, with the large number of possible robot/foundation/valence variations (36 in total), the aim in this study was not to compare responses across these variations. Instead, the aim was to broadly and inductively describe people's understandings of robots-as-PMPS, covering a range of robot morphologies, moral modules, and moral valences. Finally, participants completed items for individual moral values and reflections on their answers (data not analyzed here).

Stimulus Robots

To ensure that extracted patterns represent people's reactions to robots, broadly, stimulus-robot morphologies were varied: anthropomorphic, zoomorphic, or mechanomorphic (**Figure 1**). The anthropomorphic robot exhibited human properties: Robothespian with InYaFace projection head (Engineered Arts, United Kingdom), using the Pris female face and Heather American-English female voice. A recording of the Heather voice was dubbed over the other two robots so that variation among robots was limited to visual properties. The zoomorphic robot was spider-like: the six-legged Hexa (Vincross, China). The mechanomorphic robot exhibited overtly machine-like properties (i.e., not innately human or animal). A review of the ABOT database (Phillips et al., 2018) robots with 1–10%

TABLE 1 | Perceived moral patiency elicitations, by foundation.

Foundation	Scenario: Ray goes out into the world as it usually does, and then encounters a human who treats it. . .	Elicitation: What would it look like for Ray to be . . .
Care	with care	treated with care—with kindness and gentleness for its physical mental, or emotional well-being?
Harm	harmfully	treated harmfully—with harshness and disregard for its physical, mental or emotional well-being
Fairness	with fairness	treated with fairness—where the human acts in a way that supports justice for Ray, treats Ray equally, and/or allows Ray to have rights or opportunities equal to those of others?
Unfairness	with unfairness	treated unfairly—where the human acts in a way that was unjust to Ray, that doesn't allow Ray the same opportunities as others, and/or cheats Ray out of some kind of right or potential benefit?
Loyalty	with loyalty	treated with loyalty—where the human acts in a way that is faithful, devoted, or otherwise dedicated to Ray?
Betrayal	with betrayal	treated with betrayal—where the human acts in a way that is unfaithful, traitorous, or otherwise disloyal to Ray?
Authority	as an authority	treated like an authority—where the human acts in a way that is subordinate, obedient, or otherwise respectful to Ray's higher status, leadership, or expertise?
Subversion	as something to be undermined	subverted—where the human acts in a way that undermines Ray by being disobedient, overbearing, sabotaging, or otherwise disrespectful to Ray's authority, status, or knowledge?
Purity	as something to be kept pure	treated as something to be kept pure—where the human acts in a way helps Ray to keep clean, innocent, or otherwise fresh and uncontaminated?
Degradation	as something that should be contaminated	treated like something that can be corrupted—where the human acts in a way that degrades, spoils, or otherwise pollutes or contaminates Ray?
Liberty	as something that deserves liberty	treated with liberty—where the human acts in a way that helps Ray to be free, independent, or to otherwise determine what it wants to do?
Oppression	with oppression	treated with oppression—where the human acts in a way that enslaves, constrains, or otherwise limits Ray's independence?

human-likeness often featured a single base or wheels, a single eye (if any), and a square, round, or arm-like shape with a shiny and/or white surface. On these criteria the mechanomorphic robot was one-eyed, stationary, monolithic: the Bac configuration of Clicbot (KEYi Tech, China). In all cases, the robots called themselves "Ray."

The robots delivered (*via* pre-recorded video) an identical self-introduction message with semantic gesturing. This message included the robot's name, emphasized that it exists in the world similarly to and differently from humans, that a sophisticated body and computing equipment allows it to participate in various worldly activities, and that when it goes out into the world it seeks out things that are special and interesting. This script was designed to convey conditions by which people could possibly interpret moral patiency: the possibility (but not necessity) of general patiency, the ability to encounter human agents in the world, and a recognition that there are both good and bad phenomena. See online supplements for complete videos.

Story Elicitations

For each assigned moral foundation, participants were presented with an elicitation—an open-ended prompt that presented the scope and focus of a requested response without dictating the exact nature of how participants should respond. Each elicitation contained a label for the scenario that included a foundation-name keyword (e.g., "care"), presented an abstract scenario about Ray encountering a human, then asked what it would look like for a human to treat a robot in a specific way (each based on MFT-module descriptions; see **Table 1**). For all elicitations, participants were asked to "Please write a brief (3–5 sentence) story about a situation where a human would treat Ray in that way."

Measures

Simple metrics captured descriptive attributes of participants. Prior experience with social robots was measured using a single Likert-style item (1–7: no experience at all to extremely high experience) and liking of social robots was measured using the five-item, 7-point liking subscale of the Godspeed inventory (α = .93; Bartneck et al., 2009). A single categorical item requested self-assignment to political ideology (liberal, moderate, conservative), and demographics were drawn from Prolific's database.

Analytical Approach

Open-ended responses were subjected to inductive thematic analysis separately for each of the 12 elicitations; responses for all three robots were combined in line with the aim of identifying holistic patterns applicable to social robots, broadly. For each response set, analysis was conducted in six stages (after Braun and Clarke, 2006): Deep reading, generation of initial codes (coding unit: discernible discrete representations of human action, intent, or disposition), de-duplication of initial codes, iterative aggregation of codes into categories and then into subthemes and then into themes based on semantic similarity, checking themes for fidelity with originating data, and naming and definition of themes. Themes were identified according to keyness (utility in answering the research question) and frequency (here, mentioned in at least 10% of the coded data units; cf. Braun and Clarke, 2006). Theme frequencies varied widely across response sets given differences in how respondents addressed prompts. Misunderstandings or non-address of prompts, total rejection of a prompt's premise, and responses with no discernible human action or orientation were excluded from analysis.

TABLE 2 | Hierarchical structure of moral-patiency themes and sub-themes, by foundation.

Foundation	Theme	Subthemes
Care	Engage (*n* = 126)	Polite norms, prosocial disposition, conversation, relationship development, sharing
	Affirm (*n* = 92)	Acknowledging personhood, patient-directed conversation, deference
	Guard (*n* = 75)	Protection of body, assurance of functioning, removal from harm, recognition of risk/vulnerability
Harm	Attack (physical) (*n* = 131)	General physical mistreatment, direct physical aggression, indirect physical aggression, compromising bodily integrity
	Attack (verbal) (*n* = 54)	Harassment, insults, mocking, intimidation
	Objectify (*n* = 51)	Compromising personhood, diminishing agency, repurposing the body, disregarding, kicking out of the way
Fairness	Humanize (*n* = 84)	Humanistic treatment, social integration, equitable behavior/access
	Elevate (*n* = 48)	Rights advocacy, deference, preference
	Redress (*n* = 34)	Defense, protection, restoration
Unfairness	Separate (*n* = 108)	Social separation, physical separation, social mistreatment
	Compromise (*n* = 76)	Do physical harm, appropriate entitled resources, suffer undue consequence, deny assistance
	Delete (*n* = 44)	Denial of agency, ontological separation, obstruction
Loyalty	Bond (*n* = 106)	Befriend, persistently engage, egoistic attachment, ingrouping
	Protect (*n* = 91)	Privilege, protect from harm
	Serve (*n* = 77)	Maintain, assist, support purpose, deference
Betrayal	Exploit (*n* = 73)	Exploit, objectify, manipulate, supplant, schadenfreude, compel wrongdoing
	Deceive (*n* = 62)	Bait and switch, deceive, undermine
	Discard (*n* = 45)	Ostracism, abandonment, negative affect
Authority	Acquiesce (*n* = 99)	Deference, obedience
	Venerate (*n* = 91)	Adulation, respect, self-deprecation, appreciation
	Petition (*n* = 53)	Request help, benefit from superiority
Subversion	Resist (*n* = 41)	Verbal belligerence, bodily action, disabling
	Invalidate (*n* = 29)	Call into question, conspicuous invalidation, refusing authority premise
	Reject (*n* = 24)	Disobey, ignore, reject
Purity	Preserve (*n* = 123)	Safeguarding, cleaning, containing
	Manage (*n* = 101)	Manage opportunity, manage perception
	Curate (*n* = 36)	Limit problematic information, promote wholesome information
Degradation	Injure (*n* = 68)	Direct hardware injury, indirect hardware injury, defacement, infection
	Corrupt (*n* = 53)	Corrupting, abusing, hacking
	Manipulate (*n* = 29)	Inducing illegal behavior, inducing immoral behavior, impairing functions
Liberty	Cultivate (*n* = 97)	Teach, empower, facilitate
	Cede (*n* = 45)	Desist, loosen, make space
	Construct (*n* = 32)	Manifest, design, advocate
Oppression	Restrict (*n* = 66)	Constrain experience, constrain sociality, limit movement
	Diminish (*n* = 35)	Objectify, prevent self-actualization
	Force (*n* = 32)	Force labor, act against will, command

n *values are counted at the mention level; there may have been multiple mentions of discrete subthemes within individual responses.*

Analysis was guided by the sensitizing concept (Bowen, 2006) of humans' morally agentic action. Specifically, analysis primarily attended to human action or orientations (e.g., beliefs, intentions) that were explicitly described or easily inferred as directed toward the target robot. From that focus, themes took the form of present-tense verbs describing general classes of (im)moral action toward robots—that is, actions that take up robots as the benefiting or suffering patients. Because the aim of this analysis was to offer thick description of and hierarchical relations among actions manifesting robot moral patiency, it is outside the scope of this analysis to include formal coding or comparison among the robot types; such analysis is a suggested direction for future work. Complete narratives detailing the interpretive analysis process are included in the online supplements.

RESULTS

Participants reported low perceived experience with social robots ($M = 1.95$, $SD = 1.25$) and a moderately high liking of social robots in general ($M = 4.77$, $SD = 1.27$). From the thematic analysis, three key and sufficiently frequent themes were extracted for each moral-foundation valence. The hierarchical theme structure and frequencies are presented in **Table 2**, with the themes explicated below. Illustrative data excerpts integrated into these theme descriptions are presented *in italics; in some cases they have been edited for readability* (e.g., corrected spelling, removal of interjections).

Care

Engage: Engagement of the robot as a matter of positive relatedness, varying in required commitment and relational roles. At minimum, this includes civil engagement through polite norms (e.g., formalities and nonverbals: *The lady smiles and … asks how Ray is doing.*). More deeply, it may include engagement *via* generally prosocial disposition (kind, social, respectful, civil) including giving positive feedback (praise, compliment, thanks). It includes a general striking up of polite conversation (kind, respectful, civil, intelligent). At the most intimate, engagement includes development of relationships (e.g., *tries to become emotionally connected*) and the sharing of experiences (spending time together—usually walking).

Affirm: Acknowledgment and affirmation of the robot's existence, identity, consciousness, intelligence, and/or equivalence to humans. This may be enacted through deference to the robot as a legitimate agent, especially in exhibiting intentions or desires to provide aid or support while allowing the robot to retain agency over the nature of the care (e.g., *would also not touch Ray without Ray's permission*); affirmation is often initiated by the human actor inquiring as to the robot's operational or emotional well-being. This theme also includes conversation grounded in the robot's unique interests—its welfare, experiences, or personal life, characterized by listening, deference, interest, and understanding the robot as an individual (e.g., *[A child] spends the rest of the day assembling small rock piles, and is delighted when Ray considers them good and beautiful*).

Guard: Proactive and reactive address of the robot's physical and operational well-being. Proactive physical care includes anticipating or recognizing risk, either embodied (malfunctions, vulnerabilities: *keep it safe or protect its processors and mechanisms*) or environmental (hazards, obstacles, harmful humans); it may also include gentle handling or regular review of maintenance requirements. Reactive address includes attending to known bodily issues (cleaning, drying, fixing, or attempting human analogs like feeding) or removal from problematic situations [placing in safe location, freeing when stuck: *asks if (Ray) is lost and pulls out their smart phone. They get directions to the location and accompany Ray to the destination*…].

Harm

Attack (physical): Actions directly or indirectly impacting the robot's bodily well-being and integrity. Direct attacks are those committed by the human's own body or extending instruments: breaking, smashing, kicking, hitting, vandalizing, or degrading (e.g., *attempts to expose its wires and cause it to malfunction*). Indirect attacks are those in which an uncontrolled instrument is used to inflict harm, such as throwing an object. Attacks also include situations in which the human puts the robot into a harmful or compromising situation: throwing into the trash or fire, sending the robot's body to the ground, or stressing the limits of its functioning/capacity (e.g., *puts a bag… on her to stress her motor functions*).

Attack (verbal): Use of speech to denigrate the robot for the specific purpose of inflicting psychological or social harm. These actions include general verbal abuse (picking on, rudeness, *saying derogatory or insulting things*), making fun of (mocking: *call it names, like can opener or microwave*), and intimidation (threatening, yelling). Verbal harassment may be augmented by physical aggression but is principally enacted through language.

Objectify: General diminishing of and disregard for the robot as an agent—not necessarily to cause harm (because the robot is not seen capable of experiencing harm) but to serve the opinions, interests, or convenience of the human. These include diminishing its person-status (questioning legitimacy/realness, rejecting autonomy, reducing to object: *"You are not real"* and *"You are just a robot."*) and disregarding capacities for opinions or feelings. It may also include diminishing of agency through incapacitation (silencing, disabling, immobilization) or impedance (disrupting or blocking), removal when seen as a barrier, or repurposing its parts for practical or financial gain (e.g., *takes parts off of Ray and sells them at the junk yard for scrap metal*).

Fairness

Humanize: Engaging attitudes, behaviors, or practices grounded in a belief that robots are not—but should be—treated the same as humans. This is achieved through attempts to socially integrate robots by offering invitations, conversing and engaging according to human norms, participating in joint activities, and by otherwise engaging in human-equivalent behaviors, job assignments, benefits, being-status recognition, and civility. In short, *[g]iving Ray the same respect they would give to anyone they encounter on the street*. It also includes offering robots informational or environmental resources when they may be at a disadvantage compared to humans (e.g., *helping the robot claim its spot* when people are cutting in line), sometimes on the grounds that they are not well-equipped to independently handle human contexts.

Elevate: Enacting behaviors or practices that amplify or advance the interests of the robot, principally by privileging or deferring to the needs, desires, thoughts, and feelings of the robot. These behaviors are sometimes a matter of implicitly or explicitly recognizing that the robot is deprivileged by default and must be actively privileged as a matter of equity. Sometimes, the robot is elevated through recognition of the robot's specialness, and thus given preference over humans as a matter of its inherent superiority or vulnerability. Elevation also includes human advocacy of robots' entitlement to equal and/or constitutional rights and amplification of robots' subjectivity (e.g., *This human advocates for robots like Ray by creating and signing petitions in favor of legislation protecting robots from exploitation.*).

Redress: Acting in ways that aim to restore fairness in the wake of potential or actual harm because of some vulnerability, *will go out of the way to save Ray* from humans. This includes protection against threat or other potential harm, defense against enacted attempts at harm. It also includes restoration of physical or resource losses following some committed injustice, for instance a *human may feel as though they may cut it in line because [Ray is] not real* … [another human] *may step up and defend Ray*.

Unfairness

Separate: Disallowing social interaction *via* separation from others or through mistreatment that makes it undesirable; this separation is implicitly characterized as denial of common rights to relate to other social agents. Separation may be a social parceling-out, in which the robot is ostracized through rejection or ignoring, prevented from participating in relational activities (*e.g., a human could still choose only humans to form the team*), or more fundamentally silenced (disallowed a voice). It may also take the form of social antagonism, where a human thinks, feels, or more actively evangelizes the robot's non-belonging or non-participation.

This denial of engagement may also be enacted through physical separations, where the robot is refused access to public spaces (for instance, *via a sign saying no AI allowed*), segregated from humans, excluded from social events, or more overtly removed or relocated. More indirect forms of separation come in the mistreatment of robots when they do engage, including general meanness, diminishing status or reputation (e.g., insults, discrediting, rejection of abilities, vilification).

Compromise: Diminished security through the intentional risking of well-being or resource access, suggesting that it is unfair for an entity to be made intentionally at-risk. Compromising well-being included direct or indirect physical harm, degradation or destruction, prevention of power access, emotional harm, or the violation of harm protections (e.g., *in a way that would violate whatever warranty was given*). The robot may also suffer an appropriation, theft, or other unjust loss of resources to which it is otherwise entitled, including practical (e.g., losing one's spot in line), material (e.g., theft, trickery), and information (e.g., preventing access) resources. Compromising may also include the refusal of justified assistance such as reasonably expected services and information. Security may also be compromised when the robot takes actions in accordance with rules and norms (doing the right thing, helping, following laws) but nonetheless suffers undue consequence—as in using an open power outlet to charge only to be met with a human who *tells her to beat it and maybe even picks her up and plugs in his phone*. Similarly, this theme includes suffering what may be called "injury to insult"—there is some indirect harm suffered because of a more basic denial of expected resources, as when denial of a bus ride prevents it from achieving a goal.

Delete: Voiding social or operational value as an agent. A robot's agency is denied when it is treated as void of autonomy or sovereignty—that it is unalive, has no feelings or thoughts, may be reasonably forced or coerced in alignment with human desires, or is a non-person, effectively deleting its relevance. The underlying assumption is that one may reasonably expect to be recognized as having inherent value. Whereas separation features parceling-out by time and activity, devaluing features a more fundamental separation of robots into an ontological category based on non-humanness—most often as inferior, substandard, or excluded (e.g., upon arriving at a deli as part of a foursome, the hostess says, *Excuse me, but don't you mean 3?*). As a non-person, the robot may be subject to obstructive action by humans: impeding progress or achievement, denying opportunities to work or learn and, thus, preventing advancement in knowledge, skill, and experience (e.g., *a boss decides not to hire Ray due to how different Ray is because it is a machine*).

Loyalty

Bond: Forming bonds, most prominently through the adoption of the robot as a friend or companion, often developing feelings, perspective-taking, general liking, or desires to be near. Bonds may be formed through persistent engagement, as people interact regularly or intensely with robots toward active seeking-out of company, persistent copresence, interdependence (e.g., *addicted to the companionship*), or more generally existing in long-term relationships. Sometimes humans may work to actively ingroup the robot, integrating it into social circles like families, engaging it as they would humans or pets (*give him a squirt of oil, as I would give a dog a treat*), working to teach them how to exist with humans, or advocating for inclusion. However, sometimes the bond is an egoistic attachment, where loyalty serves humans' interests by affirming of self (e.g., pride in the association) or by fulfilling some desire or need (e.g., *understand that it is a useful resource*).

Protect: Protecting the robot from harm by humans or circumstance through proactive interventions (e.g., chaperoning) or instruction (e.g., threat identification), through defense against some negative action from humans (e.g., *standing up for Ray*), or through more general ensuring of safety and care (e.g., watching out for, relief of burdens). Protection may also come in the form of privileging the robot, elevating it above some kinds of harm. This privileging may be in relation to other technologies (e.g., *would not want to trade Ray for another robot*) or to humans (seeing humans as inferior, being willing to favor over humans).

Serve: Active or passive accommodations for the robot, generally performed consistently over time. Maintenance was most common, as the assurance of continued operation by performing upkeep of the robot's technical needs (daily *nice cleaning with an alcohol wipe*) or supporting avoidance of known operational risks (e.g., providing shelter). Other active service included helping when the robot cannot otherwise accomplish a task, helping out of an unfortunate situation (e.g., being overwhelmed), or helping to learn about the world and advance skills, knowledge, and experiences. Service can also come in the form of supporting the robot's purpose or mission by more passively working to understand it and/or evangelizing and participating in its purposeful action—even to the point of being *willing to turn against other humans in support of Ray*. Most passively, deferent service included listening, asking, obeying, following, and fearing, as well as exhibiting one's dedication through promises and making oneself vulnerable to the robot.

Betrayal

Exploit: Treatment of the robot as a tool for achieving humans' own ends. This included general taking-advantage-of (e.g., *hurt people or commit crime* on humans' behalf) and which was sometimes exacerbated by blaming and harming after having received some benefit from the robot. Objectifying practices underscored exploitation by treating it as a tool or as property—an object that has worth but may also be disregarded when the human saw fit. Sometimes exploitation manifested *schadenfreude*, or a relishing or thrill in seeing the robot degraded, harmed, or antagonized (as with *record*[ing] *a video of Ray getting blasted to bits by an oncoming vehicle* and uploading to social media). Achieving these ends could be enacted through manipulation (e.g., threatening, confusing), compelling some wrongdoing (by convincing, forcing), or even by supplanting the robot by replacing it with or demoting it to other superior robots.

Deceive: Performing bait-and-switch manipulations in three forms: bait and refuse (promise without delivery), bait and reverse (offering or giving and then reneging or taking-away), and bait and compromise (making some invitation or promise and then endangering or harming). Although this was sometimes in the interest of exploitation, it was most often characterized as a form of autotelic betrayal rather than as use-for-means, as when Ray (serving as a bartender) is *duped by the human with a recipe for a drink that no one would find appealing*. Also prevalent were overt descriptions of deception (trickery, obfuscation, or *manipulat*[ion of] *facts to confuse Ray*) to create harm or disadvantage, as well as undermining through sabotage or otherwise setting up to fail by, for instance, first praising and then discrediting the robot.

Discard: Social and functional rebuffing in three forms: ostracism, abandonment, negative affect. Ostracism included forms of domination, exclusion, discrimination, and control as a sort of girdling and parceling-out of the robot from social contexts. Most commonly there was advocacy for institutional control, especially in the restraining, arresting, and policing of robots, however at the individual also worked to maneuver exclusions (for instance, through the videorecording of harm and posting to social media). Also prevalent was discarding through abandonment: purposeful stranding (as with *a disaster where they tell Ray they will come back for her*), ignoring, or neglect. There were also relational abandonments such as cheating (as would a spouse) or breaching trust by *inform* [ing] *others of Ray's secrets without Ray's consent*. Discarding also included more passive holding of negative affect—principally mistrust, resentment, and disdain.

Authority

Acquiesce: Acquiescence in two forms: deference to the robot and more submissive obedience to it. Deference included humans making themselves second to the robot in conversation (not speaking until spoken to, listening intently), in physical presence (e.g., giving way when it needed to pass), in matters of intelligence (*defers to Ray's superior knowledge*), and in its relative role (as it may enjoy a higher status as a supervisor or cultural figure). Obedience comprised hard-and-fast compliance (*complies with Ray's order*), especially as the robot may take up gatekeeper roles in mediating access to monetary, spatial, or information resources. For obedience in supervisory relationships, humans may work to be industrious toward timely and effective performance for the robot and apologize or make appeals for transgressions. People may also engage norms for obedience associated with a robot's legal or institutional role, as when it functions as part of the police, military, or government.

Venerate: Active or passive adulation—worshiping, idolizing, or loving, or more public evangelism of the robot's worth while following it as a leader. This following was sometimes expressed in the trope of welcoming *robot overlords*. Adulation also included fear in the sense that the robot's intelligence, embodied strength, or social power could have consequences (so one should *make sure to get on Ray's good side*) as well as faith, belief, or confidence in the robot's aims and methods. People may offer due respect, especially in terms of being polite and kind. Commonly, veneration took the form of twin comparisons: recognition of the robot's superior knowledge and abilities (feeling awe, fascination, envy, admiration, thankfulness) and self-deprecation while understanding humans' inferiority (vulnerability, lower intelligence).

Petition: Requesting assistance, especially as the robot functions in a service or high-performance capacity. This most often including asking for help (e.g., *come to Ray for practical advice*, instructions, directions, opinions) and especially in relation to its higher intelligence (expertise, memory) and/or higher performative capacity (e.g., being *a bad ass robot*, being *a wealth of information without interjecting an emotional tone*). This was generally self-interested petitioning, to achieve some goal or derive some benefit—even to the point of becoming dependent on its help.

Subversion

Resist: General working-against the robot by verbal or physical action in reaction to its implicit or explicit authority. It includes verbal belligerence, *insulting, or otherwise disrespectful*. Most frequent was the physical resistance associated with disabling the robot by impairing its hardware (e.g., *dismantling part of her*) or manipulating its software (e.g., *attempt to hack it*) such that it cannot function properly. Other forms of physical resistance come in humans using bodies against it, such as cutting in line (i.e., demoting the robot in a queue) or trying to *undermine Ray by becoming physically abusive*.

Invalidate: Action that erodes the underpinnings of the robot's authority. Most common were forms of conspicuous invalidation like critiquing, or mocking, or creating situations where it would look incompetent—most specifically undermining analysis (e.g., *changes some of the data… to trick Ray and change his prior, and accurate, analysis*). Key to this invalidation is that there is some audience for the action where the subversive sentiment of the actor may be seen by and ideally spread to others. Invalidation also includes thoughts or actions that call into question the robot's authority (e.g., being *very distrustful*) and very often refusing the premise of the robot's authority altogether. The most specifically rejected premises are that information creates power and that *cold logic* can govern human affairs.

Reject: Dismissal of the robot's information, direction, or action, overwhelmingly by ignoring the robot: disregarding its instructions, suggestions, attempts to intervene, warnings (even at the human's own peril). Rejection frequently manifested as disobedience (the robot says one thing and the human does another). Sometimes this spurning came in more overt forms, including *insisting on speaking to an actual person*.

Purity

Preserve: Keeping the robot's body whole and intact by safeguarding (protecting, warning, instructing helping, defending) from harmful events, agents, situations, or spaces. Often this included keeping the robot clean and uncontaminated by performing maintenance, removing contaminants, or otherwise promoting tidy or even pristine states. It alternately includes prevention of harm by *keep*[ing] *it in a secluded place* because *being anywhere in the world would contaminate Ray*: out of harm's way, redirected from harmful spaces, or through (in) voluntary containment. Containment most often included bringing it into one's home or putting it into a box, case, or secret place where it cannot be exposed to harm or harmful agents act upon it.

Manage: Controlling external influences on the robot by supervising the robot's opportunity for certain experiences and guiding interpretations of experiences. Regarding opportunity, humans may limit it to good experiences and interactions (*to show the world in a good light*), prevent negative or problematic experiences (e.g., *going into a seedier portion of the city*), or eliminate the opportunity to have experiences altogether. Regarding interpretations, experiences encountered may be framed in a positive light to protect the robot from understanding them fully and being influenced by that negativity—hiding, distracting, disregarding, or candy-coating the world's harsh realities.

Curate: Overseeing the robot's access to and engagement with information, especially by limiting problematic information and promoting wholesome information. Often this limiting and promotion is performed through curation of media exposure (wholesome: *Hallmark Channel*, *the best cartoons*, *Leave it to Beaver*, *love songs*; problematic: *internet, commercials, book*[s] *about murder*). More generally, it includes prevention from learning about problematics of humanity (*abuse, poverty, crime, pollution, bias, violence, death*), selection of clean conversation topics (*weather, favorite colors*), and the general embargoing of harsh or profane language.

Degradation

Injure: Decaying a robot's physical form directly or indirectly, but always purposefully. This degradation includes injury to the body proper by hitting, breaking, torturing, rending, dissembling, destroying, melting, environmental exposure, or spilling of substances onto it. Although not always explicitly said, the sentiment throughout these mentions was an intention to break down the body—especially into substructures that were less offensive, threatening, or unappealing. Degradation of bodies also included defacement, most often to *vandalize Ray with graffiti, as they might do to a bathroom wall*. Some suggested that the body could be degraded through infection: sneezing, spitting, touching, urinating, or (maskless) coughing.

Corrupt: Distorting information inherent to the robot and its functioning or the information it is exposed to through experiences. Most common were references to perverting the robot through exposures to corrupt information *via* impure experiences, media content, or communication. Sometimes this effort was *trying to get Ray to say something offensive to make people laugh… to swear or be racist*. This is especially so for verbal abuse (insulting, degrading, harsh address)—distinct from speech inherent to harm in that it included clear sentiments of tearing down the robot using words (e.g., *telling it how it's unnatural*). Information corruption also took the form of hacking, with human actors *trying to de-program or re-program Ray to do something it is not intended to do or designed to do*.

Manipulate: Influencing behaviors, often for the human's own benefit (e.g., entertainment, revenge), to induce generally or specifically illegal and/or immoral behavior. Illegal behavior included efforts to *rob money* or information, but also included *us*[ing] *Ray for covert surveillance*, casting illegal votes, and polluting. Immoral behavior comprised bad, unkind, unethical actions. Manipulation also included the intentional impairment of the robot's functions to prevent understanding of its experiences, movement, or environmental sensing (e.g., *cover her face so she couldn't see… She would not understand what had happened and would either report a malfunction in her camera or keep trying until her battery died.*).

Liberty

Cultivate: Creating the conditions for liberty within the robot itself. Cultivation included empowerment by first discovering the robot's subjectivity (held or desired purpose, opinions, desires, plans, feelings, interest, barriers, wishes, thus *respected for having his own free will*), then facilitating the realization of that independent subjectivity. Facilitation came in the form of helping to overcome barriers or reach goals, or protecting against threats to those goals. Sometimes this came in being a sidekick: accompanying the robot or even deferring one's own activities and interests. Alternately, humans may cultivate independence by inspiring the robot through discussions of the future and of possibilities, or teaching it specific skills for independence (e.g., *practice reasoning with her* or *show them how to be independent*). Teaching may also include explanations of notions of freedom, independence, and rights, or even working to convince the robot to value those principles: *to advocate for Ray and guide them through discovering independence and making decisions… similar to raising a child.*

Cede: Degrees of diminished interference in the robot's affairs. At the lowest degree, this included efforts to *give it a looser leash* in the form of constrained freedoms, most often giving options and allowing choices from those options (e.g., *choose which path to follow*) or more liberally to allow for it to make choices within rules, laws, reason, or moral boundaries. More often it was a general leaving-be: not interfering, bothering, meddling, or even interacting with the robot as a means of allowing it to deal with its affairs unfettered such that humans would not *impede Ray's ability to determine what it wants to do*. Making space was another form of ceding human control over the robot by giving it space to move without obstruction, a space of its own apart from human interference, or even adapting existing spaces to be well-suited to the robot.

Construct: Overt actions to directly manifest native or emergent liberty. Most commonly, humans liberate robots by altering hardware or software to ensure freedom from control or by commanding it into freedom. It also included engineers or computer scientists designing independence into the robot *via* its programming (e.g., decision-making, resilience) or hardware (e.g., agile legs for self-sufficiency)—for example, *Ray's engineers could give it freedom by designing flexible limbs to help it maneuver in different environments, and programming to help it make its own interpretations about input it receives.*

Oppression

Restrict: Constraining movement through imprisonment and/or immobilization. Imprisonment compromises the enclosure of robots into a box, room, or cage, generally for purposes of asserting control over it or secreting it away. Immobilization is the limiting or prevention of movement by confining it to specific spaces, boundaries, or distances, such as allowing it to roam only inside a home or tethering it to a human. Restriction also extended to limiting social interaction (*shut Ray in a room and not allow it to interact with other beings*) or outright silencing (*not allowing it to speak freely*). More generally, humans may delimit experiences and subjective growth by disallowing access to the world or to meaningful experiences, or by restricting independent thought. Sometimes this occurs through the disabling of specific abilities (GPS for navigation, sensors for sight), but more often was associated with imprisonment or confinement and characterized as blocking input or stimulation.

Diminish: Systematic depreciation of the robot to a mere thing—a toy, object, piece of property, or something expendable—often for the human's own benefit. Generally, taking away a higher status was associated with subjugation, as with *see*[ing] *Ray as inferior and not allowing Ray to move freely in society*. As something that could or should have some higher status, diminishing also included the active prevention of self-actualization (often by restriction described above). Humans may prevent a robot from realizing its purpose or, most often, assign it a diminished role that disregards its abilities (e.g., being a store greeter means that *his immense knowledge base would be completely wasted*).

Force: Compelling into labor, especially by human command or coercion, usually by threat of destruction or deactivation if the robot does not comply. This laboring was sometimes characterized as enslavement (relegated to *human uses from the day it was created*) and was often described in superlatives—the robot does *everything*, *all the time*, and that is its only role. Commands or physical manipulation may also be used to force a robot to act against its will or without regard to the outcome, including situations in which a robot may be ordered to effectively work itself to death, *until his internal parts were no longer operational*.

DISCUSSION

This study elicited stories about the ways that people see robots as viable moral patients through the lens of humans' (im)moral actions, extracting themes that both comport with and deviate from conceptualizations of human moral patiency. Although the primary aim of this work was descriptive, patterns across described moral upholdings and violations also illuminate the importance of ontological categorization and how people may make meaning across category boundaries.

Perceived Moral Patiency and Ontological Categorization

Robots may be perceived as moral patients in ways that reflect both benefit and suffering. Moral benefit across the foundations often took the form of humans working to integrate the robot into human society (social engagement, affirmed personhood, humanization, status elevation, bonding through in-grouping). This pattern of beneficence-as-integration signals that PMP may rest on recognitions that social robots are "othered" (Kim and Kim, 2013)—set apart from humans by their origin, tool status, dependencies, lack of emotion, and different intelligence (Guzman, 2020). This othering has implications for how people morally engage robots (Edwards, 2018): In supporting robots' well-being or preventing their suffering, humans would maximize similarities or minimize differences from humans. Although most upholding themes could be reasonably applied to human PMPs, some relied on robots' differences from humans. Most notably, upholding authority included a human agent benefiting from a robot's authority (egoistic or utilitarian rather than altruistic drives; cf. Singer, 2011), manifesting liberty by design (grounded in robots' made-not-born origins; Mayor, 2018), and upholding purity by curating inputs so as not to contaminate the outputs (indirect impacts on humans-as-users; cf. Friedman, 2020). Moral suffering found robots to be generally diminished and set apart from humans (objectified, separated, devalued, discarded, rejected, invalidated), such that moral victimization seems to be meaningfully linked to perceived *it*-ness (rather than *who*-ness) of robots; this inanimacy corresponds with seeing robots as property (see Edwards, 2018). Perhaps unsurprisingly, these patterns were especially prevalent within harm, unfairness, and oppression—i.e., violations of the "individualizing" foundations that, when upheld, emphasize the rights of individuals (*versus* loyalty, authority, and purity, which emphasize social cohesion; Haidt and Graham, 2007).

Altogether, findings are interpreted to suggest that ascribing moral patiency to robots is largely a function of how one engages social robots' liminal ontology. That is, social robots are of a kind that exhibits both human and machine characteristics such that they do not map clearly to either category (Kahn et al., 2011); these ontological hybridity may activate overlapping mental models (see Banks et al., 2021) where one must determine whether robots are more like humans or more like machines. Indeed, there are many themes that could easily apply to humans (e.g., guard, attack, redress, compromise, serve, deceive, petition, resist, manage), suggesting a privileging of human-likeness. Importantly, there are also themes that explicitly call out robots' ontological liminality through upholding/violating juxtapositions: in care/harm (engaging or affirming personhood *versus* overt objectification), fairness/unfairness (humanizing/elevating status *versus* separating from humans or deleting existence), and loyalty/betrayal (bonding as a humanlike friend or discarding as an unneeded object). Authority/subversion, purity/degradation, and liberty/oppression themes do not exhibit this overt ontological-category engagement or separation, but still hint at the sentiment in respectively invalidating the premise of robot authority, degradation through a hacking-into, and assumptions that robots are innately oppressed and must be actively freed. Thus, PMP may be shaped by categorical presuppositions (Coeckelbergh, 2018; Edwards, 2018): Moral treatment of robots is shaped by applying norms and assumptions associated with humans, and immoral treatment is shaped by rejecting humanizing norms and/or embracing those for machines. In other words, mental models for the robot-as-PMP include some degree of acceptance or rejection of its personhood and mode of existence. This is, of course, not a surprising inference as it is argued in ethics and philosophy domains (e.g., Floridi and Sanders, 2004; Danaher, 2020; Gunkel and Wales, 2021), but here it has been empirically derived (see also Guzman, 2020). Other work details humans' tendencies to draw boundaries around components of the world, where a "moral circle" is a boundary that separates those that are deserving of moral consideration and those that do not (Singer, 2011, p. 120). In framing the decision of who belongs inside *versus* outside that circle, people who take an exclusionary mindset have larger (i.e., more inclusive) circles, while those that focus on who to include have smaller circles (Laham, 2009). Individual engagement of robots-as-PMP, then, likely depends on the framing of an encounter, as well as a host of other personological and intergroup variables.

The Devil(ishness) in the Details

That people were able to imagine situations in which robots are patients to humans' (im)moral actions is itself important in that it reveals the potential for robots to socially (rather than merely functionally) integrated into human social spheres. That is, people could imagine human-machine relations where the robot meaningfully experienced repercussion of human action, which requires the inferencing of a robot's internal (i.e., mental or embodied) states (see Banks, 2020a) and of its integration with human moral norms (cf. Malle and Scheutz, 2014). Importantly, as the aim of this working was to describe the nature of and conditions for robot PMP, nuance is always lost in the extraction of broad patterns. The finer details of the elicited narratives—though not rising to the criteria for themes—illuminate some hints as to how moral patiency is similar but perceptually distinct for robots, compared to humans. To care for a robot may include replications of human care but may also include perspective-taking resulting in recognition that robot needs are different, such as crafting things the robot would find interesting. To be unfair to a robot includes treating it in a way that violates its warranty, negatively impacts its intellectual development, or prevents it from accessing power resources. To be loyal includes allying with the robot over humans, including supporting an A.I. uprising, and disloyalty includes embarrassing it by preying on misunderstanding of human norms. Recognizing robot authority was less about deferring to human institutions like tradition, religion, or government and more about acknowledging human inferiority in speed, accuracy, and precision. Subverting robot authority, in turn, often relied on trickery (such as corrupting data inputs) to undermine that performative superiority. Despite purity's biological conceptualization in MFT, degradation took on violations of bodily and operational integrity: defacement, induced glitching, and forced illegal behaviors. Liberty was sometimes seen as manifested by humans through hacking or design, while oppression was sometimes about obstructing information access.

The presence of certain concepts in robot-as-PMP narratives is notable: information, intelligence, standards, and operation. These do align with common concepts in mental models for robots more generally (Banks, 2020b) but may serve particular functions in ascribing moral status. Specifically, because they are relevant to both humans' and robots' operation, they may serve as boundary objects—ideas that are concrete enough to have a specific meaning, but plastic enough to be interpreted differently and adapted across groups (Star and Griesemer, 1989). As such, they may serve a translational function by "developing and maintaining coherence across intersecting social worlds" (p. 393), fostering cooperative common ground without necessarily requiring exactly similar interpretations (Bechky, 2003). These plastic concepts may function as boundary objects that facilitate metaphorical thinking (see Koskinen, 2005)—and that thinking may allow people to consider robots' moral status without necessarily drawing on human criteria, especially as the metaphorical boundary objects are clarified through use over time. For instance, the notion that it is unfair to violate a robot's warranty can be likened to a violation of human rights to healthcare. The "warranty" has particular-but-parallel meanings in humans' and machines' ostensible social worlds: guarantee of corrective address of bodily integrity issues. Those types of objects may be a bridge to developing interventions, either supporting or suppressing perceptions of robots' PMP. Although it is outside the scope of this work for advocating for or against the ascription of moral patiency (i.e., the can/should dimensions), identification of these objects as bridging concepts (i.e., "transcendental language;" Coeckelbergh, 2018) serves as a fruitful direction for future research into how moral status may (not) be ascribed.

Limitations and Future Research

This study is subject to common limitations in interpretive research (idiosyncrasies of the researcher lens, selection of a single solution from among all possible interpretations, interpretation of participant's meanings without probing). These should be addressed through replication of the work, along with empirical testing of the claims regarding the role of perceived ontological liminality in the ascription to and operation of moral-patient status. The specific elicitations and robots used to garner PMP stories may have influenced the nature of the stories told, and other characterizations of moral modules or stimulus robots could elicit different kinds of stories. Further, this study accounted for perceptions of only three robot morphologies (when there are many variations on the three classes and other classes altogether) and offered only a brief and decontextualized introductory video. As the three stimulus robots were engaged here to ensure that PMP themes were extracted from stories about a range of robots, the present work did not examine differences across those stimuli. Future research should establish the extent to which identified themes are applicable across robots with different characteristics and in different contexts—especially when robots are co-present rather than presented in a mediated fashion. Moreover, because moral status emerges in relation to temporal and cultural norms (Coeckelbergh, 2020), this work can only be taken as a starting point—limited to the late-2020 United States zeitgeist—and patterns in mental models for robots as PMPs are likely to shift as both technology, culture, and corresponding dispositions change. Nonetheless, the themes identified or the types of human action inducing perceived moral patiency of robots—are a useful foundation for future work on the antecedents, dynamics, and effects of PMP in human-machine interaction. Specifically, future work should draw on identified themes as a framework for the construction of stimuli and measurement that reflect humans' innate understandings of the potentials for robot PMP.

CONCLUSION

People imagine often-rich scenarios in which robots are moral patients to humans' (im)moral actions—from affirming robots' personhood as acts of care to objectifying them as acts of oppression. When people perceive social robots to be moral patients, they draw from intersecting notions of moral action and subjection inherent to both human life and machine operations. From that frame, ascription of moral-patient status to robots may reflect dispositions toward ontological separations between human and machine—breaking down separations toward moral upholding, embracing separations toward moral violations, and sometimes for both moral valences engaging an entanglements of "is" and "is-not" human. Identifying concepts that have concrete-yet-plastic meaning in both human life and machine operations may be a vehicle for understanding the ways in which people do and do not expand circles of moral concern to include social machines.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/5pdnc/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Texas Tech University Human Research Protection Program. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-19-1-0006.

ACKNOWLEDGMENTS

The author gratefully acknowledges Kristina McCravey for technical and logistical support and Patrick Caster for stimulus video production.

REFERENCES

Agrawal, S., and Williams, M. (2017). Robot Authority and Human Obedience: A Study of Human Behaviour Using a Robot Security Guard. *HRI'17 Companion*, 57–58. ACM.

Anderson, D. L. (2013). "Machine Intentionality, the Moral Status of Machines, and the Composition Problem," in *Philosophy and Theory of Artificial Intelligence*. Editor V. C. Mueller (Heidelberg, Germany: Springer), 321–333. doi:10.1007/978-3-642-31674-6_24

Aroyo, A. M., Kyohei, T., Koyama, T., Takahashi, H., Rea, F., Scuitti, A., and Sandini, G. (2018). "Will People Morally Crack under the Authority of a Famous Wicked Robot?," in International Symposium on Robot and Human Interactive Communication, New Delhi, India (IEEE), 35–42.

Asaro, P. M. (2006). What Should We Want from a Robot Ethic?. *Int. Rev. Inf. Ethics* 6, 9–16. doi:10.4324/9781003074991

Banks, J., Edwards, A. P., and Westerman, D. (2021). The Space between: Nature and Machine Heuristics in Evaluations of Organisms, Cyborgs, and Robots. *Cyberpsychology, Behav. Soc. Networking* 24 324–331. doi:10.1089/cyber.2020.0165

Banks, J. (2020a). Good Robots, Bad Robots: Morally Valenced Behavior Effects on Perceived Mind, Morality, and Trust. *Int. J. Soc. Robotics*. [online before print] Retrieved from. doi:10.1007/s12369-020-00692-3

Banks, J. (2020b). Optimus Primed: Media Cultivation of Robot Mental Models and Social Judgments. *Front. Robotics AI* 7, 62. doi:10.3389/frobt.2020.00062

Bartneck, C., and Keijsers, M. (2020). The Morality of Abusing a Robot. *Paladyn. J. Behav. Robotics* 11 (1), 271–283. doi:10.1515/pjbr-2020-0017

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *Int. J. Soc. Robotics* 1 (1), 71–81. doi:10.1007/s12369-008-0001-3

Baum, F. L. (1904). *The Marvelous Land of Oz*. Chicago, IL: Tae Reilly Britton. doi:10.1007/978-3-642-47360-9

Bechky, B. A. (2003). Sharing Meaning across Occupational Communities: The Transformation of Understanding on a Production Floor. *Organ. Sci.* 14 (3), 312–330. doi:10.1287/orsc.14.3.312.15162

Bowen, G. A. (2006). Grounded Theory and Sensitizing Concepts. *Int. J. Qual. Methods* 5 (3), 12–23. doi:10.1177/160940690600500304

Bradwell, H. L., Johnson, C. W., Lee, J., Winnington, R., Thill, S., and Jones, R. B. (2020). Microbial Contamination and Efficacy of Disinfection Procedures of Companion Robots in Care Homes. *PLoS ONE* 15 (8), e0237069. doi:10.1371/journal.pone.0237069

Braun, V., and Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qual. Res. Psychol.* 3, 77–101. doi:10.1191/1478088706qp063oa

Bryson, J. J. (2018). Patiency Is Not a Virtue: the Design of Intelligent Systems and Systems of Ethics. *Ethics Inf. Technol.* 20 (1), 15–26. doi:10.1007/s10676-018-9448-6

Chang, W., Wang, H., Yan, G., Lu, Z., Liu, C., and Hua, C. (2021). EEG Based Functional Connectivity Analysis of Human Pain Empathy towards Humans and Robots. *Neuropsychologia* 151, 107695. doi:10.1016/j.neuropsychologia.2020.107695

Coeckelbergh, M. (2012). *Growing Moral Relations: Critique of Moral Status Ascription*. London, United Kingdom: Palgrave Macmillan. doi:10.1057/9781137025968

Coeckelbergh, M. (2010). Moral Appearances: Emotions, Robots, and Human Morality. *Ethics Inf. Technol.* 12, 235–241. doi:10.1007/s10676-010-9221-y

Coeckelbergh, M. (2020). Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking about Animals and Humans. *Minds. Machines*. doi:10.1007/s11023-020-09554-3

Coeckelbergh, M. (2018). Why Care about Robots? Empathy, Moral Standing, and the Language of Suffering. *Kairos: J. Philos. Sci.* 20 (1), 141–158. doi:10.2478/kjps-2018-0007

Connolly, J., Mocz, V., Salomons, N., Valdez, J., Tsoi, N., Scassellati, B., and Vázquez, M. (2020). "Prompting Prosocial Human Interventions in Response to Robot Mistreatment," in International Conference on Human-Robot Interaction, Cambridge, United Kingdom (New York, NY: ACM), 211–220.

Cox-George, C., and Bewley, S. (2018). I, Sex Robot: The Health Implications of the Sex Robot Industry. *BMJ Sex. Reprod. Health* 44, 161–164. doi:10.1136/bmjsrh-2017-200012

Craik, K. (1943). *The Nature of Exploration*. Cambridge, United Kingdom: Cambridge University Press.

Danaher, J. (2017). Robotic Rape and Robotic Child Sexual Abuse: Should They Be Criminalised? *Criminal L. Philos.* 11, 71–95. doi:10.1007/s11572-014-9362-x

Danaher, J. (2020). Welcoming Robots into the Moral circle: A Defence of Ethical Behaviourism. *Sci. Eng. Ethics* 26, 2023–2049. doi:10.1007/s11948-019-00119-x

Darling, K., Nandy, P., and Breazeal, C. (2015). "Empathic Concern and the Effect of Stories in Human-Robot Interaction," in International Symposium on Robot and Human Interactive Communication, Kobe, Japan (IEEE), 770–775 .

Davenport, D. (2014). Moral Mechanisms. *Philos. Technol.* 27, 47–60. doi:10.1007/s13347-013-0147-2

de Graaf, M. M., and Malle, B. F. (2019). "People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences," in International Conference on Human-Robot Interaction, Daegu, Korea (IEEE), 239–248.

Dubal, S., Foucher, A., Jouvent, R., and Nadel, J. (2011). Human Brain Spots Emotion in Non Humanoid Robots. *Soc. Cogn. Affective Neurosci.* 6 (1), 90–97. doi:10.1093/scan/nsq019

Edwards, A. P. (2018). "Animals, Humans, and Machines: Interactive Implications of Ontological Classification," in *Human-machine Communication: Rethinking Communication, Technology, and Ourselves*. Editor A. L. Guzman (New York, NY: Peter Lang), 29–50.

Eyssel, F., and Kuchenbrandt, D. (2012). Social Categorization of Social Robots: Anthropomorphism as a Function of Robot Group Membership. *Br. J. Soc. Psychol.* 51 (4), 724–731. doi:10.1111/j.2044-8309.2011.02082.x

Floridi, L., and Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14 (3), 349–379. doi:10.1023/b:mind.0000035461.63578.9d

Fraune, M. R., Šabanović, S., and Smith, E. R. (2017). "Teammates First: Favoring Ingroup Robots over Outgroup Humans," in International Symposium on Robot and Human Interactive Communication (IEEE), 1432–1437.

Friedman, C. (2020). "Human-Robot Moral Relations: Human Interactants as Moral Patients of Their Own Agential Moral Actions towards Robots," in *Artificial Intelligence Research*. Editor A. Gerber (Cham, Switzerland: Springer), 3–20. doi:10.1007/978-3-030-66151-9_1

Gary, H. E. (2014). Adults' Attributions of Psychological agency, Credit, and Fairness to a Humanoid Social Robot. [Dissertation]Retrieved from https://digital.lib.washington.edu/researchworks/handle/1773/27568?show=full Accessed by May 10, 2021.

Gong, T., and Cai, Z. (2008). Tri-tier Immune System in Anti-virus and Software Fault Diagnosis of mobile Immune Robot Based on normal Model. *J. Intell. Robot Syst.* 51, 187–201. doi:10.1007/s10846-007-9186-1

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the Moral Domain. *J. Personal. Soc. Psychol.* 101 (2), 366–385. doi:10.1037/a0021847

Gray, K., and Wegner, D. M. (2009). Moral Typecasting: Divergent Perceptions of Moral Agents and Moral Patients. *J. Personal. Soc. Psychol.* 96 (3), 505–520. doi:10.1037/a0013748

Gray, K., Young, L., and Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychol. Inq.* 23 (2), 101–124. doi:10.1080/1047840x.2012.651387

Grodzinsky, F., Wolf, M. J., and Miller, K. (2019). "Applying a Social-Relational Model to Explore the Curious Case of hitchBOT," in *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*. Editors D. Berkich and M. d'Alfonso (Cham, Switzerland: Springer), 311–323. doi:10.1007/978-3-030-01800-9_17

Gunkel, D. J. (2018). The Other Question: Can and Should Robots Have Rights?. *Ethics Inf. Technol.* 20 (2), 87–99. doi:10.1007/s10676-017-9442-4

Gunkel, D. J., and Wales, J. J. (2021). Debate: what Is Personhood in the Age of AI?. *AI Soc.* [online before print]. doi:10.1007/s00146-020-01129-1

Guzman, A. (2020). Ontological Boundaries between Humans and Computers and the Implications for Human-Machine Communication. *Human. Machine. Communication.* 1, 37–54. doi:10.30658/hmc.1.3

Haidt, J., and Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals May Not Recognize. *Soc. Just Res.* 20, 98–116. doi:10.1007/s11211-007-0034-z

Iyer, R., Koleva, S., Graham, J., Ditto, P., and Haidt, J. (2012). Understanding Libertarian Morality: The Psychological Dispositions of Self-Identified Libertarians. *PLoS One* 7 (8), e42366. doi:10.1371/journal.pone.0042366

Kahn, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., et al. (2012). "Robovie, You'll Have to Go into the Closet Now": Children's Social and Moral Relationships with a Humanoid Robot. *Develop. Psychol.* 48 (2), 303–314. doi:10.1037/a0027033

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Shen, S., Gary, H. E., and Ruckert, J. H. (2015). "Will People Keep the Secret of a Humanoid Robot?: Psychological Intimacy in HRI," in Annual International Conference on Human-Robot Interaction, Portland, OR (New York, NY: ACM), 173–180.

Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., and Gill, B. T. (2011). "The New Ontological Category Hypothesis in Human-Robot Interaction," in International Conference on Human-Robot Interaction (New York, NY: ACM), 159–160.

Keijsers, M., and Bartneck, C. (2018). "Mindless Robots Get Bullied," in International Conference on Human-Robot Interaction, Chicago, IL (New York, NY: . ACM), 205–214.

Kim, M.-S., and Kim, E.-J. (2013). Humanoid Robots as "The Cultural Other": Are We Able to Love Our Creations?. *AI Soc.* 28 (3), 309–318. doi:10.1007/s00146-012-0397-z

Koskinen, K. U. (2005). Metaphoric Boundary Objects as Co-ordinating Mechanisms in the Knowledge Sharing of Innovation Processes. *Euro Jrnl of Inn Mnagmnt* 8 (3), 323–335. doi:10.1108/14601060510610180

Laham, S. M. (2009). Expanding the Moral circle: Inclusion and Exclusion Mindsets and the circle of Moral Regard. *J. Exp. Soc. Psychol.* 45 (1), 250–253. doi:10.1016/j.jesp.2008.08.012

Lima, G., Kim, C., Ryu, S., Jeon, C., and Cha, M. (2020). Collecting the Public Perception of AI and Robot Rights. *Proc. ACM Human-Computer Interaction* 4 (CSCW2), 135. doi:10.1145/3415206

Maggi, G., Dell'Aquila, E., Cuccineiello, I., and Rossi, S. (2020). "Cheating with a Socially Assistive Robot? A Matter of Personality," in HRI'20 Companion, Cambridge, United Kingdom, 352–354. ACM.

Malle, B. F., and Scheutz, M. (2014). "Moral Competence in Robots," in Proceedings of the International Symposium on Ethics in Science, Technology and Engineering, Chicago, IL (IEEE), 1–6 .

Mayor, A. (2018). *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Berlin, Germany: De Gruyter. doi:10.1515/9780691185446

Milman, A., Tasci, A., and Zhang, T. (2020). Perceived Robotic Server Qualities and Functions Explaining Customer Loyalty in the Theme Park Context. *Int. J. Contemporary. Hospitatily. Manage.* 32, 3895–3923. doi:10.1108/ijchm-06-2020-0597

Nijssen, S. R., Hyselaar, E., Müller, B. C., and Bosse, T. (2020). Do we Take a Robot's Needs into Account? the Effect of Humanization on Prosocial Considerations toward Other Human Beings and Robots. *Cyberpsychology, Behavior, and Social Networking* 24 (5), 332–336. doi:10.1089/cyber.2020.0035

Phillips, E., Zhao, X., Ullman, D., and Malle, B. F. (2018). "What Is Human-like?: Decomposing Robot Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database," in International Conference on Human-Robot Interaction, Chicago, IL (New York, NY: ACM), 105–113.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. (2009). "How Anthropomorphism Affects Empathy toward Robots," in International Conference on Human-Robot Interaction, La Jolla, CA (New York, NY: ACM), 245–246.

Rouse, W. B., and Morris, N. M. (1985). On Looking into the Black Box: Prospects and Limits in the Search for Mental Models. *Psychol. Bull.* 100 (3), 349–363 .

Sandoval, E. B., Brandstetter, J., Obaid, M., and Bartneck, C. (2016). Reciprocity in Human-Robot Interaction: A Quantitative Approach through the Prisoner's Dilemma and the Ultimatum Game. *Int. J. Soc. Robotics* 8 (2), 303–317. doi:10.1007/s12369-015-0323-x

Sembroski, C. E., Fraune, M. R., and Šabanović, S. (2017). "He Said, She Said, it Said: Effects of Robot Group Membership and Human Authority on People's Willingness to Follow Their Instructions," in International Symposium on Robot and Human Interactive Communication (IEEE), 56–61.

Singer, P. (2011). *The Expanding circle: Ethics, Evolution, and Moral Progress*. Princeton, NJ: Princeton University Press. doi:10.1515/9781400838431

Sparrow, R. (2004). The Turing Triage Test. *Ethics Inf. Technol.* 6, 203–213. doi:10.1007/s10676-004-6491-2

Spence, P. R., Edwards, A., and Edwards, C. (2018). Attitudes, Prior Interaction, and Petitioner Credibility Predict Support for Considering the Rights of Robots. *HRI'18 Companion* (pp. 243–244). ACM.

Star, S. L., and Griesemer, J. R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Soc. Stud. Sci.* 19 (3), 387–420. doi:10.1177/030631289019003001

Sundar, S. S. (2020). Rise of Machine agency: A Framework for Studying the Psychology of Human-AI Interaction (HAII). *J. Computer-Mediated Commun.* 25 (1), zmz026. doi:10.1093/jcmc/zmz026

Sung, J.-Y., Guo, L., Grinter, R. E., and Christensen, H. I. (2007). "My Roomba Is Rambo": Intimate home Appliances," in International Conference on Ubiquitous Computing, Innsbruck, Austria (Springer), 145–162.

Trovato, G., Ham, J. R. C., Hashimoto, K., Ishii, H., and Takanishi, A. (2015). "Investigating the Effect of Relative Cultural Distance on the Acceptance of Robots," in International Conference on Social Robotics, Paris, France (Springer), 664–673. doi:10.1007/978-3-319-25554-5_66

Ward, A. F., Olsen, A. S., and Wegner, D. M. (2013). The Harm-Made Mind. *Psychol. Sci.* 24 (8), 1437–1445. doi:10.1177/0956797612472343

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer (AE) declared a past co-authorship with the author (JB) to the handling Editor.

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 07 July 2021
doi: 10.3389/frobt.2021.712427

Whether to Save a Robot or a Human: On the Ethical and Legal Limits of Protections for Robots

*Kamil Mamak**

Department of Criminal Law, Jagiellonian University, Kraków, Poland

Proponents of welcoming robots into the moral circle have presented various approaches to moral patiency under which determining the moral status of robots seems possible. However, even if we recognize robots as having moral standing, how should we situate them in the hierarchy of values? In particular, who should be sacrificed in a moral dilemma–a human or a robot? This paper answers this question with reference to the most popular approaches to moral patiency. However, the conclusions of a survey on moral patiency do not consider another important factor, namely the law. For now, the hierarchy of values is set by law, and we must take that law into consideration when making decisions. I demonstrate that current legal systems prioritize human beings and even force the active protection of humans. Recent studies have suggested that people would hesitate to sacrifice robots in order to save humans, yet doing so could be a crime. This hesitancy is associated with the anthropomorphization of robots, which are becoming more human-like. Robots' increasing similarity to humans could therefore lead to the endangerment of humans and the criminal responsibility of others. I propose two recommendations in terms of robot design to ensure the supremacy of human life over that of humanoid robots.

Keywords: moral patiency, moral circle, robot rights, moral dilemma, trolley problem

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Sven Nyholm,
Utrecht University, Netherlands
Joshua Gellers,
University of North Florida,
United States

****Correspondence:***
Kamil Mamak
kamil.mamak@uj.edu.pl

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

***Received:** 20 May 2021*
***Accepted:** 24 June 2021*
***Published:** 07 July 2021*

Citation:
Mamak K (2021) Whether to Save a Robot or a Human: On the Ethical and Legal Limits of Protections for Robots. Front. Robot. AI 8:712427. doi: 10.3389/frobt.2021.712427

INTRODUCTION

Robots are increasingly entering the social lives of humans, which raises certain questions about our mutual interaction, such as whether robots are mere tools or something more, how we should treat robots, whether we owe robots anything, and whether robots should have rights. In recent years, increased academic attention has been paid to such issues, and many important publications have been published on these themes (cf. Balkin 2015; Darling 2016; Gunkel 2018b; Pietrzykowski 2018; Turner 2018; Abbott 2020; Bennett and Daly 2020; Gellers 2020; Nyholm 2020; Smith 2021). Schröder stated that "controversies about the moral and legal status of robots and of humanoid robots in particular are among the top debates in recent practical philosophy and legal theory" (Schröder 2020, 191). The discussion of robots' possession of rights is strongly connected with deliberation on their moral status, another of the principal topics considered in the ethics of artificial intelligence (Gordon and Nyholm 2021). A few review works concerning such issues have recently been published (Schröder 2020; Gordon and Pasvenskiene 2021; Harris and Anthis 2021).

In this paper, I focus on the limits of the protection of robots by answering the question of who should be saved–human or robot. Some people have indicated that they would hesitate to sacrifice robots to save humans. Nielsen et al. examined how the anthropomorphization of robots impacts the decisions of humans in a moral dilemma when there is a need to sacrifice one entity to save another. The authors' results indicate that "when people attribute affective capacities to robots, they become less likely to sacrifice this robot to save a group of human beings" (Nijssen et al., 2019, 53). These results are alarming from the perspectives of both ethics and law. Current legal systems take the stance that human life is at the top of protected values. Furthermore, not saving humans in a situation in which there is the possibility of doing so could be considered a crime.

As robots are becoming increasingly human-like, this issue will continue to gain importance over time. The following question thus emerges: Should we act in order to maintain human life as the most valuable from the legal perspective? For example, if we accept that human life should always be at the top of hierarchies of value, perhaps manufacturers should be forced to mark robots such that they can be easily differentiated from humans in emergencies. In unforeseen traffic accidents, drivers only have seconds to decide what to do and what they can avoid. Robot drivers and human drivers should know that robots should be sacrificed in collisions involving both humans and robots. From another perspective, we should ask whether robots have any properties that make them equal to humans with regard to legal protections, such as a human-like intelligence, and whether we could in fact decide that robots should be granted more protection than humans. I respond to all of these issues in this paper, which is structured as follows.

I start by considering the issue of rights for robots and presenting popular ways of ascribing moral patiency to robots. I then explore conflict situations between the lives of robots and those of humans on the basis of the presented approaches. The subsequent section is devoted to the contemporary hierarchy of values set by law; here, I demonstrate that a person who hesitates to sacrifice a robot could be considered to have committed a crime. Finally, I offer recommendations for modifying the design of robots to mitigate the described risks and present the conclusion of this study.

RIGHTS FOR ROBOTS?

Could a robot have rights? The short answer to this question is "yes". Law is a social technology (Fairfield 2021), and we can, in theory, do whatever we want with it. According to a popular anecdote, Caligula made his beloved horse Incitatus a consul. Whether true or not (and it seems not; Barrett 2015, 289), this anecdote illustrates that someone who has the power to create the law can theoretically do almost whatever they want. The law is a flexible tool, and if there is a need, it can be used for different purposes. Gellers, for example, noted that ships had formal legal status in history because there was such a need (Gellers 2020). Hence, there is no theoretical obstacle to granting rights to robots.

More demanding is the "should" question, which is tied to the issue of the moral standing of robots. If robots were welcomed into the moral circle, we could expect that human interactions with robots would be impacted by their possession of moral status. Some scholars categorically argue that robots should not be granted moral status (cf. Bryson 2010; Birhane and van Dijk 2020), but there is also a significant body of literature that claims otherwise. I briefly present four approaches to determining the moral patiency of robots: properties-based, indirect duties, relational ethics, and environmental ethics.

The most widely accepted approach to granting moral status to robots is based on what a robot "is". To decide whether an entity is qualified to enter the moral circle, we must know its ontology. If that ontology contains the qualities that we believe are important, we accept that the entity is in the moral circle. In some approaches, a quality or set of qualities is sufficient to resolve the moral status of robots. Other approaches discuss properties such as sentience, intelligence, or consciousness (cf. Floridi and Sanders 2004; Sparrow 2004; Himma 2009; Levy 2009; Hildt 2019; Kingwell 2020; Mosakas 2020; Gibert and Martin 2021; Véliz 2021). Thus, if a robot can feel pain or is self-aware, then we should incorporate it into the group of entities that possess moral status. An approach based on properties seems a useful tool by which to grant moral status in theory, but in reality presents several issues. First, there is no consensus as to which quality/qualities should be sufficient for moral consideration, as different authors have identified different qualities on which to ground moral patiency. Second, there is no consensus as to what human qualities are; we still do not know what it means to be conscious, self-aware, or intelligent (cf. Umbrello and Sorgner 2019). Third, as Gunkel wrote, the basis of moral status on qualities serves as a way of postponing the discussion (Gunkel 2018b). Fourth, as Coeckelbergh observed, there are epistemological limitations (Coeckelbergh 2010, 212), such as how to know whether a robot is feeling pain (cf. Dennett 1978; Bishop 2009; Adamo 2016). We already struggle to determine the inner states of other human beings; robots could be much harder to "read".

Danaher proposed an interesting response to the epistemological problem through the theory of ethical behaviorism, "[. . .] which holds that robots can have significant moral status if they are roughly performatively equivalent to other entities that have significant moral status" (Danaher 2020, 2023). Danaher did not focus on what robots are, but rather on how they perform in everyday life (i.e., the observable aspect of their functioning). If robots cross the performative threshold of entities that have moral status, we should treat them as such entities. Some scholars have criticized ethical behaviorism (Nyholm 2020; Smids 2020), and I describe one issue created by this theory in a later section. However, ethical behaviorism is the most practical response to the lack of knowledge concerning the qualities of entities with which we are interacting–if we believe that qualities matter.

The second popular approach to moral patiency is grounded on the Kantian theory of indirect duties toward animals. Kant believed that animals do not have (direct) moral status, but that humans should treat them well regardless. He claimed that

> if a man has his dog shot, because it can no longer earn a living for him, he is by no means in breach of any duty to the dog, since the latter is incapable of judgment, but he damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind. Lest he extinguish such qualities, he

must already practice a similar kindliness toward animals; for a person who already displays such cruelty to animals is also no less hardened toward men. (Kant 1997, 212).

Proponents of this theory liken robots to animals in order to advocate for granting moral status to robots and thereby preserving our own humanity. One of the proponents of this approach is Kate Darling, who developed the analogy of robots to animals (Darling 2016). Darling suggested that robots are new animals and that we should consider how humans previously resolved issues in our relationships with animals to prepare for our existence alongside robots (Darling 2021). Smith also developed a Kantian approach, which advocates treating robots as moral patients to prevent their dehumanizing use and to protect the dignity of humans (Smith 2021). Coeckelbergh connected this approach to the relational turn, which I briefly discuss below (Coeckelbergh 2020b).

The relational turn in roboethics is largely associated with two authors, Coeckelbergh and Gunkel, who did not limit their deliberations to robots (cf. Coeckelbergh and Gunkel 2014). In their view, the moral patiency of robots is not grounded on robots' ontological properties, but, crucially, on the relations between robots and humans (Coeckelbergh 2010). In this approach, ethics precedes ontology and is usually the opposite (Gunkel 2018b). As Gunkel noted, "[. . .] the question of social and moral status does not necessarily depend on what the other is in its essence but on how she/he/it (and the pronoun that comes to be deployed in this situation is not immaterial) supervenes before us and how we decide, in "the face of the other" (to use Levinasian terminology), to respond" (Gunkel 2018a, 96). Coeckelbergh claimed that "we could argue that, [. . .], the status of AIs will be ascribed by human beings and will depend on how they will be embedded in our social life, in language, and in human culture" (Coeckelbergh 2020a, 59).

The last approach that I want to mention here is environmental ethics. This approach is neither fully distinct from the concepts presented earlier (e.g., the approach based on a Kantian view of animals) nor a homogeneous concept. Different strands in environmental ethics differ in their response to questions concerning how humans should relate to the environment and non-human entities and how to situate humans among them (cf. Brennan and Lo 2021). However, in his book on rights for robots, Gellers embedded the issue of robots within the concept of environmental ethics, suggesting that determining the moral standing of robots could be a "side effect" of discussion of the moral status of nature and its elements. Gellers advocated for a critical environmental ethics approach according to which the idea of recognition of robots' rights (and those of other non-natural entities) is related to epistemic pluralism (Gellers 2020). This approach may, for example, be focused on the harmony between the elements of nature, with one such element being technological artifacts, including robots. If we grant moral standing to trees (cf. Stone 2010), why not to robots? This environmental approach is also supported by the religious beliefs of non-Western cultures, which are discussed in depth in Gellers' book.

ROBOTS' RIGHT TO LIFE

Moral standing may be granted to robots on many different grounds. Possessing moral consideration is the basis for possessing rights (Danaher 2020). However, the question remains which rights are to be possessed. Accepting the notion that it is possible for robots to possess rights says little about the content of such rights; this is another issue that requires further deliberation (cf. Graaf et al., 2021). Humans enjoy various types of rights, ranging from the right to privacy to the right to free expression and the right to holidays. Some of these rights are transferable to robots, while others are not. Furthermore, robots could potentially have specific rights resulting from their distinct ontology. This paper, however, is not the appropriate place to expand on this issue; instead, I limit my deliberations to the concept of what could be called the "right to life."

The right to life is one of the basic rights that might be derived from the acceptance of robots' moral standing; here, I use "might," because it is not obvious (see Lima et al., 2020, 135: 6). I would like to note two objections: First, assessing whether robots exist or not alone is problematic. Indeed, such determinations are problematic for humans as well. The criteria used to determine death are legally and ethically unclear and have changed during the course of history, for example, from the irreversible loss of heart and lung function to the death of the brain (cf. Belkin 2014; De Georgia and Michael, 2014). There are still occasional protests concerning whether we should turn off life-support apparatuses, even in cases where brain death has been confirmed. Assessing whether a robot no longer exists could be even more problematic. Is a robot "dead" when all of its data are stored online, but the physical body is destroyed? Second, it is not clear that the right to existence is a basic right. If we use the example of animals, we could say that some animals are in the moral circle in terms of animal rights; still, it is possible to kill even these animals for certain purposes, such as for food or clothes. From a legal perspective, it is possible for a farmer who breeds animals to kill them legally and to be punished for cruelty to the very same animals he kills. For the purpose of further deliberation, I ground the notion of the right to life in the meaning that any breach of that right will destroy a robot completely.

What should be noted is that the deliberations on a robot's "right to life" do not mean that "robot rights" are some kind of an extension of human rights. In this particular instance there is just a similarity to the concept of the right to life (which itself belongs to the domain of human rights). The set of possible robot rights is different from the set of human rights (cf. Gunkel 2020).

Different elements of our social and biological world-such as e.g., corporations, animals and nature-have already been determined to possess certain rights in different places around the word (see more on that Gellers 2020). The discussions about particular rights of [certain kinds of] robots should be treated similarly–that is: as discussions on rights of non-humans. What is more, robots (like pets or farm animals) are someone's property, and this characteristic makes them legal objects and not legal subjects. Legal subjects are legal persons - both natural and artificial (i.e., corporations)-and legal personhood is associated with a wider scope of legal rights (more on the concept of legal personhood: Kurki 2019). Having in mind this division into legal objects and subjects the treatment of robot rights as an extension of human rights seems even more inaccurate.

It is one issue to claim that robots have moral standing; how we situate robots in the hierarchy of values is another issue. Now I turn to how best to resolve the dilemma of whether the lives of robots should be more or less valued than the lives of humans, or, in other words, who we should save according to the previously presented approaches.

Let us assume that properties such as sentience and intelligence are not binary concepts, but a spectrum. We could, on that ground, say that different entities are situated at different points on such scales. In his book *Superintelligence*, Bostrom situated different organisms in this way regarding their intelligence (Bostrom 2016). Simpler organisms are lower on the scale, while human beings are at the top. This thinking allows us to assert that, for now, human beings are at the top of the scale, which justifies their privileged position over other inhabitants of our planet. However, what if robots were to exceed humans in terms of those properties that we believe provide the basis for moral standing? Should we recognize their superiority over us and, for example, prioritize them in a dilemmatic situation? My deliberations here are extremely speculative due to the plurality of philosophical concepts involved and the problem of epistemological limitations. However, a question like this could arise at some point, especially in the context of the priorities that the law assigns to human beings. We must think about how we want to organize our world with regard to entities that are situated at different points on the scale in relation to human beings.

There are three potential answers to the question about prioritization. If robots possess qualities that correspond to qualities of entities that are lower on the hierarchy of values (e.g., robots with insect-like intelligence), we should prioritize humans. A more complicated answer results in the case of entities that are, more or less, the same as humans. Bearing in mind how difficult it may prove to determine what is "like" a human, we can imagine that robots may be like us. In this scenario, it seems appropriate that we should treat robots as equal to human beings. In (Putman, 1964) observed that the materials used in the construction of a robot should not matter; what should matter is the qualities the robot possesses (1964). Prioritizing human beings could be seen as discrimination based on the materials used to build an entity. In this thinking, the question of prioritization is unanswerable; it would be similar to asking whether we should prefer older people to younger people or men to women. Such an *a priori* decision could be seen as discrimination and thus be forbidden by law. The most controversial answer would result if robots outperform humans in the qualities that we consider to be a source of moral standing. It cannot be excluded that priority should b given to robots, and Sparrow defends such a position (Sparrow 2004).

The approach based on Kantian indirect duties is the easiest means to answer the prioritization problem. Kant believed that animals do not have moral status; therefore, robots also do not have (direct) moral status. Humans, however, do have such status. Thus, in conflict situations, we should save human beings.

In contrast, the relational approach is the most unclear in regards to resolving the prioritization problem. This approach is focused on the relations of human beings with robots, not on robots' ontology. On the one hand, the relational approach says little about how to deal with a conflict situation. On the other hand, this approach is, in a sense, anthropocentric. The relations that ground moral standing originate from humans; human relations are the starting point for ethical decisions. From that perspective, human beings will take precedence over any other entities with whom humans have relations. Gunkel adopted the relational view proposed by Levinas on the grounds of roboethics and also admitted that Levinas made an anthropocentric interpretation of his own works (Gunkel 2018a, 97). However, during the recent workshop "Rabbits & Robots: Debating the Rights of Animals & Artificial Intelligences" organized by the Cambridge Centre for Animal Rights Law (Cambridge Centre for Animal Rights Law 2021), Coeckelbergh suggested that the relational approach is anthropocentric, but epistemically, and not necessarily morally. Nevertheless, even taking this clarification into account, it is still unclear whether it is permissible to sacrifice humans to save robots.

Environmental ethics is not homogenous, and there are different possible answers under this approach to the prioritization question. In the anthropocentric view of the environment, priority is given to human beings. The modern version of anthropocentrism is called "enlightened anthropocentrism" or "prudential anthropocentrism" (Brennan and Lo 2021). These views regarding the environment are also connected with the Kantian view presented above and similarly resolve the conflict situation, namely by answering that human beings should be saved. According to anthropocentric environmental ethics, there is no priority granted to robots as non-human entities, nor to other non-humans. However, it is also possible to arrive at the opposite answer on the grounds of non-anthropocentric environmental ethics, such as bio-or ecocentrism. For example, in eccentric environmental ethics, humans do not enjoy priority over other species, as the ecosystem is considered as a whole. Describing the deep ecology movement, which could be seen as ecocentric, Naess stated that it rejects "the man-in-environment image in favour of the relational, total-field image" (Naess 1973, 95). The non-anthropocentric view was also advocated by Gellers (Gellers 2020). Gellers' critical environmental ethic is ecocentric and holist, positing that all vulnerable entities present in an open ecology are radically equal. His approach takes inspiration from non-Western and indigenous worldviews. In the context of non-anthropocentric approaches, it is worth mentioning a case from 2016, when a Cincinnati Zoo worker killed a gorilla to protect a three-year-old who had fallen into the gorilla's enclosure (Panagiotarakou 2016). In that case, environmental ethicists were not certain that the zookeeper should have killed the gorilla (cf. Bein and McRae 2020), indicating that this perspective is open to the possibility that non-human beings have priority. If the destruction or killing of non-human entities would do irreparable damage to nature and the harmony between its elements, then a human being could be sacrificed.

In sum, who should we save, the human or the robot? The answer is most ambiguous under the properties-based approach. In some scenarios, the properties-based approach indicates that the priority should be given to humans, but, in others, prioritizing humans could be considered an act of discrimination. It is even possible to imagine that priority should be given to robots if their qualities outperform those that we believe to be the basis of moral standing. In an approach based on Kantian indirect duties, the answer is clearer: We should save human beings, as they are the entities with direct moral status. In the relational approach, the priority is also (probably) given to humans as the source of the relations. Finally, on the basis of environmental ethics, the answer depends on the initial starting point. The anthropocentric approach prioritizes human beings, but the answer is more unclear in relation to non-anthropocentric views, according to which preference may be given, in some cases, to non-human lives.

Although it is beyond this paper's scope, which is dedicated to the conflict between humans and robots, another intriguing version of the prioritization problem could arise when we raise similar questions in the context of a dilemma involving animals and robots (cf. Wilks et al., 2021). There is already a growing body of literature looking at the interactions among animals and robots (cf. Butail et al., 2014; Romano et al., 2019).

SAVING ROBOTS INSTEAD OF HUMANS IS A CRIME

Previous deliberations concerning ethics have been normative in nature, focusing on how humans should behave and starting with different ethical assumptions. These deliberations are useful for discussing how humans should organize our mutual life with robots in the future. All of the previously introduced approaches are theoretically possible to adopt, with some obstacles. Indeed, some of them are already part of the social order, such as non-anthropocentric environmental approaches in some Native American tribal lands (see Gellers 2020). However, it is difficult to imagine that these approaches would be easily universalized for translation from one jurisdiction to the next, for example, into Western systems.

The current answer to the question of who should be sacrificed between humans and robots is connected to the hierarchy of values embedded in legal systems. Hesitation to sacrifice robots in order to save humans, as exhibited in the research of Nijssen et al. (2019), is highly problematic from the perspective of contemporary law, and such behavior could even be a crime. The remainder of this section focuses on this issue.

The law is human-centered, and in case of dilemmas–human life vs. non-human life, there is almost no doubt that human life is favored. The right to life and physical security is the most basic claim of every human being (Ashworth 1975, 282). According to the modern understanding of human rights, the individual human being is put in the center as the goal and the end, and the right to life is a fundamental human right (Ziebertz and Zaccaria 2019b). There is a legal obligation to protect life, and any exemptions are highly controversial, such as abortion, killing in self-defense, euthanasia, and the death penalty (see on those issues, cf. Ziebertz and Zaccaria 2019a; Fletcher, 1978). Even a cursory legal assessment reveals that the right to life of a human being is highly protected at the international, regional, and national levels. Many laws declare humans' "right to life," often citing the United Nations Universal Declaration of Human Rights from 1948, which states in Article three that "everyone has the right to life, liberty and security of person". For a prioritization dilemma, a second document is even more informative: The European Convention on Human Rights (ECHR). According to Article 2,

1. Everyone's right to life shall be protected by law. No one shall be deprived of his life intentionally save in the execution of a sentence of a court following his conviction of a crime for which this penalty is provided by law.
2. Deprivation of life shall not be regarded as inflicted in contravention of this Article when it results from the use of force which is no more than absolutely necessary: 1) in defence of any person from unlawful violence; 2) in order to effect a lawful arrest or to prevent the escape of a person lawfully detained; 3) in action lawfully.

According to this provision, there is no possibility of deprivation of human life in order to save non-human entities. Clause (a) of Article 2.2 states that deprivation of life is allowed under some circumstances if it is necessary to protect the life of another human being.

People should not only not take human life, but sometimes are obliged–under a threat of punishment–to actively protect human life. I now briefly discuss this obligation. Criminal law is one of the branches of law most resistant to harmonization, and some important features of criminal responsibility are not derived solely from a single provision. I illustrate the resulting problems using an example from a specific real-world system, namely that of Poland. Polish criminal law includes a crime called "failure to render aid," which is useful to examine the problem of hesitation to sacrifice robots to save humans. As stated in the Polish Criminal Code, Article 162, this crime is defined as follows:

§ 1. Whoever does not provide assistance to a person being in an immediate danger of loss of life or sustaining a grievous bodily harm, even though he could have provided it without exposing himself or another person to a danger of loss of life or a danger of sustaining a grievous bodily harm, is subject to the penalty of deprivation of liberty for up to 3 years.

§ 2. Whoever does not provide assistance that requires a medical procedure, or in a situation where a prompt assistance can be provided by an institution or a person responsible for providing such assistance, does not commit a crime. (Wróbel et al., 2014).

Failure to render aid is a specific type of crime. Crimes usually concern behaviors that are not permitted, such as theft, murder, and rape. However, the legal system can also punish individuals for not doing something that it believes to be desirable. The system literally forces people to do something and, if they do not, threatens punishment. Because punishment for not doing something is an unusual case, it is used only in a limited set of examples, such as to regulate the actions taken when another human being is in a life-threatening situation. The legal system takes the view that if another human is in danger, a bystander cannot just look on, but must take action to help. The criminalization of failure to render aid is justified by one of the most commonly accepted moral norms–the need to help a person whose life or health is seriously endangered (Zoll [in:] Wróbel and Zoll 2017).

A few issues connected with the crime of failure to render aid require explanation. The first is obvious: The agent who can expect help is a human being rather than any other entity, animal, or non-living artifact. Only failure to help human beings is punishable by law. The law does not oblige people–under threat of punishment – to actively save animals, trees, or artifacts if there is a threat to their existence, even if they have high material or cultural value. Simply put, it is not a crime to watch and not help when an animal is dying or a tree is falling. One might regard such an act as morally corrupt, but it does not constitute a crime.

The second issue requires more explanation and is connected with the clause in the provision that reads "without exposing himself or another person to a danger of loss of life or a danger of sustaining a grievous bodily harm." One could regard this clause as confirmation that the law does not require heroism: The obligation to act has limitations, and individuals are allowed to do nothing if there is a serious threat to themselves or to another human being. The crucial issue from the perspective of this paper is that the excuse not to help–and not to be liable for a crime–concerns the state of danger in which a human being, not an animal or any other artifact, finds themself. Individuals are obliged to do everything possible, including "sacrificing" non-human entities. Considering an illustrative example is helpful here. We can imagine that there is a person who is on fire, and a witness is wearing an expensive coat that could be used as a rescue tool. Nothing else nearby could be similarly used. The witness is obliged to use that coat to rescue the other person, even if doing so means that the coat will be destroyed. A dynamic situation such as this requires immediate reaction, which means the witness is obliged to react. If the witness does not act, he or she is committing a crime. The witness is obliged to act in the same way in terms of sacrificing animals and robots. In the case noted above of the gorilla at the Cincinnati Zoo, the zookeeper behaved correctly by saving the human child and killing the animal. The legal system is quite straightforward on this matter: The human being has a greater value, and if the zookeeper had not done what he did, he could be criminally liable for failure to render aid. Similarly, in a conflict between human life and robot "life", failure to sacrifice the robot would constitute a crime.

The important issue as concerns the crime of failure to render aid–which results not from the description of the crime, but from the general rules of criminal responsibility–is intent to commit it (Wróbel and Zoll 2014). In the common law criminal literature, intent is associated with the concept of *mens rea* (cf. Lewna 2018; Zontek 2018). Criminal intent means, in part, that the perpetrator is aware of all elements of a crime. In the case described in this paper, it means that a witness must be aware that another person is in a life-threatening situation and that they–or any other human–will not be threatened by providing help. For example, no crime will be committed if a person is lying on a bench in a park having a heart attack and needs medical intervention if the witnesses are not aware that the person needs help. A further example would be a witness who observes a child who is drowning. The witness knows that the child will die without help, but the witness cannot swim and is afraid that he or she will also die if he or she gives help. The witness is not aware that the water is 1 m deep, and there is no real threat. In this situation, no crime will be committed if the witness thinks that helping exposes him or herself to danger. According to criminal law, it is important what a person is thinking during the act under evaluation. If someone or something is deceiving a person, it will be considered. If, for example, a person thinks that he or she is interacting with humans, but is really interacting with robots (or vice versa), it could be crucial for determining criminal responsibility. If a person attacks a robot, thinking that they are attacking a human, that person could still be sentenced for a criminal attempt to attack a human, even if there was no human involved.

Hence, if a robot resembles a human and a person thinks that the robot is human and that not helping another human will save that human-like robot from danger, the person does not commit a crime. This, too, is an important notion. Lack of criminal responsibility does not mean that the situation is without difficult legal implications. The law can fail to achieve the goal of protecting humans in danger, which reveals the practical issue with Danaher's ethical behaviorism. Danaher wrote about moral rights, not legal ones. However, in implementing his position within the scope of the law, a problem emerges. Danaher proposed the "the rule of actions," which holds that we should treat robots like the entities they mimic (having in mind humans and animals); thus, if the entity resembles a human, we should treat it like a human. In this text, Danaher referred to the concept of the so-called philosophical zombie (cf. Kirk 2021) and argued that we should treat such entities as humans (Danaher 2020, 2029). The problem is not an objection to Danaher's argumentation, which is coherent, but it demonstrates that this kind of thinking could have consequences that may be contradicted by the legal system, reflecting the gradation of values that places human life, over the lives of entities that look like humans, at the top. The problem of human-like robots is not purely abstract. There are examples of such robots, such as the robotic copy of Hiroshi Ishiguro or Sophia the robot.

Returning to the research of Nijssen et al. (2019), their dilemmas were structured in the same logic: "A group of people is in danger of dying or getting seriously injured, but they can be saved if the participant decides to perform an action that would mean sacrificing an individual agent (human, human-like robot, or machine-like robot) who would otherwise remain unharmed" (Nijssen et al., 2019, 45–46). From the perspective of the crime of failure to render aid, in every case people should sacrifice robots, and, if someone hesitated to do so in real life, they would be committing a crime.

In conclusion, robots with a human-like appearance are problematic from the perspective of the hierarchy of values embedded in legal systems. The law places the value of human life at the top of protected values. The lives of both animals and robots are worth less. In a conflict situation, we are obliged to save humans and sacrifice other entities, including robots. However, two problematic cases are possible: first, if people hesitate to sacrifice a robot, knowing that it is a robot, they commit a crime, and, second, if they hesitate to sacrifice a robot, thinking that it is a human, they do not commit a crime, but the consequence of their action (i.e., the human not being rescued) is undesirable in the legal system.

RECOMMENDATIONS

In this section, I consider the appropriate response to the fact that human-like robots could pose a danger to human life by leading people to prioritize robot life. This prioritization could be done knowingly, if a person hesitates to sacrifice robots (e.g., due to empathy toward them) or unknowingly, if a person thinks that they are prioritizing a human that is in fact a robot. The deliberation in this case is based on the assumption that we want to preserve the contemporary hierarchy of values, in which human life is at the top of the protected entities in our legal system. Bryson used the term "human-centered society" (in contrast to "artifact-centered society") (Bryson 2018). She recognized the dangers of over-attachment to robots and contended that we should respond to such dangers through design:

> We design, manufacture, own and operate robots. They are entirely our responsibility. We determine their goals and behaviour, either directly or indirectly through specifying their intelligence, or even more indirectly by specifying how they acquire their own intelligence. But at the end of every indirection lies the fact that there would be no robots on this planet if it weren't for deliberate human decisions to create them. (Bryson 2010, 65).

Bryson thus concluded that if there is a problem with the design of robots, we should change it in a way that will not cause unnecessary societal costs. In her other work, she formulated associated recommendations:

> First, robots should not have deceptive appearance—they should not fool people into thinking they are similar to empathy-deserving moral patients. Second, their AI workings should be "transparent" [. . .] The goal is that most healthy adult citizens should be able to make correctly-informed decisions about emotional and financial investment. As with fictional characters and plush toys [. . .] we should be able to both experience beneficial emotional engagement, and to maintain explicit knowledge of an artefact's lack of moral subjectivity. (Bryson 2018, 23)

Such recommendations could, in theory, provide a response to the issues discussed in this paper; however, the hope that robots will not be created to look like humans is unrealistic. Danaher, in response to such recommendations, observed that "[. . .] the drive to create robots that cross the performative threshold [. . .] will probably prove too overwhelming for any system of norms (legal or moral) to constrain" (Danaher 2020, 2046). Gunkel also commented on Bryson's recommendations by suggesting that thinking requires aestheticism, which should concern designers and users, and which he doubted is possible to enforce (Gunkel 2018a, 94). The desire to create entities that mirror humans is too strong to impose a general ban on creating robots in our own image, especially taking into consideration that robots with a human appearance are not unequivocally bad. There are dozens of areas of life in which robots that resemble humans would be beneficial, including sex robots or companion robots (cf. Di Nucci 2017; McArthur 2017; Ryland 2021). The fact that a knife can be used to commit a crime does not mean that we should ban the production of knives; they are too useful in everyday life. The same consideration applies to robots. We should minimize the potential negative outcomes from the existence of robots that mimic life rather than ban their creation, which seems to be neither possible nor sufficiently justifiable.

We should construct the world that we share with robots with consideration to how humans are. Humans tend to anthropomorphize objects: "Robots are now available in physical forms and can exhibit movements that are getting impressively more human. As a result, our brain, which has evolved to interact and understand humans, is tricked into interpreting their behavior as if it were generated by a human" (Sandini and Sciutti 2018, 7:1). Humans should take this tendency into account when discussing how to organize human–robot interactions. With regard to this topic, I offer two recommendations.

1. Humanoid robots should be easily distinguishable from humans.

People should know that they are interacting with robots. A person should be able to perceive that a robot is a robot at first glance. The fact that a robot is a robot should not be revealed only through interactions, but should also be evident from a distance. For example, robots should be easy to distinguish by drivers of cars for safety reasons, so they can be sure as to who should be sacrificed in a dilemmatic situation such as a car crash. Robots' differences from humans should be apparent to help humans make appropriate decisions in dynamic situations requiring immediate reaction. This distinction may be achieved by incorporating a particular marking element into the design of robots, such as a light or an object protruding from the head.

This recommendation could be limited to certain robots used in certain contexts—especially where there is a threat to the safety of human beings. One example would be when robots go outside of the owner's home and become a participant in traffic by crossing the street. This recommendation is comparable to the requirement that drones must not fly into certain zones, such as airports' surroundings (cf. O'Malley 2019). There are reasons that justify the limitation of usage of technologies and adaptations to prioritize safety over other features, such as the freedom of flying whatever we like or to have a robot that looks a certain way, especially if the look becomes problematic. There is no need to have such limiting features for, for example, sex robots, which are and will be used almost exclusively in an intimate environment. Forcing the producer to make them look not human-like could even destroy the experience of using such robots.

2. Robots should inform other elements of the interactive environment that they are robots.

Robots should also inform other environmental elements that they are robots, even if the robots resemble humans. This will be essential in the context of autonomous cars, among other issues. There is an ongoing discussion around the ideal infrastructure of autonomous and connected vehicles (cf. Bonnefon et al., 2020), as well as what crash algorithms should be developed or implemented (cf. Nyholm 2018). From the perspective of the assumptions made in this paper, it is clear that humans should be saved; however, a car must know that something that resembles a human is not necessarily human. Cars that will replace human drivers are in development, and robots must inform such cars that they are not humans–not only through their appearance, but also in some way that may not be perceptible to humans.

It is possible that, in some cases, robots and other elements of the digital environment will "know" that robots that look like humans are not humans even without the implementation of this recommendation. The technological environment may progress beyond our current epistemological limitations and, for example, use temperature sensors that could differentiate human from non-human. We currently base our evaluations of objects, at least from a distance, mostly on visual aspects. If something looks like a human, we have no apparatus to determine that it is not a human. However, there are features that could help recognize humans among humanoid robots. One is body temperature, which is not visible to humans, but could be visible through technology. Nevertheless, if there is no temperature sensor in the future, or if that sensor will be insufficient to distinguish humans from humanoid robots in particular cases, then the proposed recommendation could be necessary.

The aim of this recommendation could be partially achieved in another way. We should make sure that the ways we are "teaching" technologies to recognize humans as elements of the environment are based not only on appearance. In cases of human-like robots, which will be elements of our social life, relying on the visual aspect could be misleading.

The proposed recommendations will not solve all of the problems caused by the deceptive human appearance of robots; rather, humans must decide based on real data. We should also communicate to society that, for now, human life has a unique value that is protected and requires other entities to be sacrificed to save it. Simply put, we should sacrifice robots to save humans, no matter how cute and human-like those robots may be.

CONCLUSION

While the issue of robot rights may be unthinkable for some, it is nevertheless becoming an increasingly serious topic of scientific deliberation, and it is increasingly difficult to pretend that this topic is unimportant. The pressing factor is the number and sophistication of contemporary robots that increasingly resemble humans. Many issues must be resolved as soon as possible, including questions concerning how humans should treat robots. Claims that robots are mere property and should be treated as such are unsatisfactory, as our interactions–in both the research environment and in real life–demonstrate that people treat robots differently. Human relations with robots are intertwined with ethics and law.

In this paper, I have focused on the limits of the protection of robots, as illustrated by the moral dilemma of who should be saved between a human or a robot. I have discussed the issue from the perspective of various approaches of ascribing moral standing to robots and have demonstrated that prioritizing humans over robots may not always be the obvious course of action. I also explored the legal perspective, which protects the superiority of human beings as a manifestation of the hierarchy of values in legal systems. If we wish to preserve that hierarchy, we must react to the process of robots becoming more human-like. Our tendency to anthropomorphize robots could disrupt that hierarchy; in response, I have proposed recommendations that could be implemented at the level of robot design. Contemporary law is not fully ready for the coexistence of humans and human-like robots.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The publication has been supported by a grant from the Priority Research Area FutureSoc under the Strategic Programme Excellence Initiative at the Jagiellonian University.

REFERENCES

Abbott, R. (2020). *The Reasonable Robot: Artificial Intelligence and the Law*. Cambridge: Cambridge University Press. doi:10.1017/9781108631761

Adamo, S. A. (2016). Do Insects Feel Pain? A Question at the Intersection of Animal Behaviour, Philosophy and Robotics. *Anim. Behav.* 118 (August), 75–79. doi:10.1016/j.anbehav.2016.05.005

Ashworth, A. J. (1975). Self-Defence and the Right to Life. *C.L.J.* 34 (2), 282–307. doi:10.1017/s0008197300086128

Balkin, J. (2015). The Path of Robotics Law. California Law Review 6. Available at: https://digitalcommons.law.yale.edu/fss_papers/5150.

Barrett, A. A. (2015). *Caligula: The Abuse of Power*. Routledge. doi:10.4324/9781315725413

Bein, S., and McRae, J. (2020). Gorillas in the Midst (Of a Moral Conundrum). *Environ. Ethics* 42 (1), 55–72. doi:10.5840/enviroethics20204216

Belkin, Gary. (2014). *Death before Dying: History, Medicine, and Brain Death*. Oxford University Press.

Bennett, B., and Daly, A. (2020). Recognising Rights for Robots: Can We? Will We? Should We? *L. Innovation Tech.* 12 (1), 60–80. doi:10.1080/17579961.2020.1727063

Birhane, A., and van Dijk, J. 2020. "Robot Rights? Let's Talk about Human Welfare Instead." *ArXiv:2001.05046 [Cs]*, January. doi:10.1145/3375627.3375855

Bishop, M. (2009). Why Computers Can't Feel Pain. *Minds & Machines* 19 (4), 507–516. doi:10.1007/s11023-009-9173-3

Bonnefon, J-F., Černy, D., Danaher, J., Devillier, N., Johansson, V., Kovacikova, T., et al. (2020). *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*. doi:10.2777/035239

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies*. Reprint edition. Oxford, United Kingdom ; New York, NY: Oxford University Press.

Brennan, A., and Lo, Y-S. (2021). "Environmental Ethics.," in *The Stanford Encyclopedia of Philosophy*. Editor E. N. Zalta (Metaphysics Research Lab, Stanford University). Available at: https://plato.stanford.edu/archives/sum2021/entries/ethics-environmental/.

Bryson, Joanna. J. (2018). Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics. *Ethics Inf. Tech.* 20 (1), 15–26. doi:10.1007/s10676-018-9448-6

Bryson, J. J. (2010). "Robots Should Be Slaves." *Close Engagements With Artificial Companions. Key Soc. Psychol. Ethical Des. Issues* 63–74.

Butail, S., Ladu, F., Spinello, D., and Porfiri, M. (2014). Information Flow in Animal-Robot Interactions. *Entropy* 16 (3), 1315–1330. doi:10.3390/e16031315

Cambridge Centre for Animal Rights Law (2021). *Online Workshop "Rabbits and Robots: Debating the Rights of Animals and Artificial Intelligences*. Available at: https://www.youtube.com/watch?v=rUxeG26dH5Q.

Coeckelbergh, M. (2020a). *AI Ethics*. Cambridge, MA: The MIT Press.

Coeckelbergh, M., and Gunkel, D. J. (2014). Facing Animals: A Relational, Other-Oriented Approach to Moral Standing. *J. Agric. Environ. Ethics* 27 (5), 715–733. doi:10.1007/s10806-013-9486-3

Coeckelbergh, M. (2010). Robot Rights? towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12 (3), 209–221. doi:10.1007/s10676-010-9235-5

Coeckelbergh, M. (2020b). Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking about Animals and Humans. *Minds and Machines*. doi:10.1007/s11023-020-09554-3

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Sci. Eng. Ethics* 26, 2023–2049. doi:10.1007/s11948-019-00119-x

Darling, K. (2016). "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects," in *Ryan Calo, A. Michael Froomkin, and Ian Kerr*. Editors R. Law First Edition (Cheltenham, UK: Edward Elgar Pub).

Darling, K. (2021). *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. New York, NY: Henry Holt.

De Georgia, M. A., and Michael, A. (2014). History of Brain Death as Death: 1968 to the Present. *J. Crit. Care* 29 (4), 673–678. doi:10.1016/j.jcrc.2014.04.015

Dennett, D. C. (1978). Why You Can't Make a Computer that Feels Pain. *Synthese* 38 (3), 415–456. doi:10.1007/bf00486638

Di Nucci, E. (2017). *Robot Sex: Social And Ethical Implications*. John Danaher and Neil McArthur. The MIT Press. Available at: https://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262036689.001.0001/upso-9780262036689-chapter-004. doi:10.7551/mitpress/9780262036689.003.0005

Fairfield, J. A. T. (2021). *Runaway Technology: Can Law Keep up?* Cambridge: Cambridge University Press. doi:10.1017/9781108545839

Fletcher, G. P. (1978). The Right to Life. *Ga. L. Rev.* 13, 1371.

Floridi, L., and Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14 (3), 349–379. doi:10.1023/B:MIND.0000035461.63578.9d

Gellers, J. C. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. Routledge. doi:10.4324/9780429288159

Gibert, M., and Martin, D. (2021). In Search of the Moral Status of AI: Why Sentience Is a Strong Argument. *AI Soc*. doi:10.1007/s00146-021-01179-z

Gordon, J.-S., and Pasvenskiene, A. (2021). Human Rights for Robots? A Literature Review. *AI Ethics*. doi:10.1007/s43681-021-00050-7

Gordon, John-Stewart., and Nyholm, Sven. (2021). *Ethics of Artificial Intelligence | Internet Encyclopedia of Philosophy*. Available at: https://iep.utm.edu/ethic-ai/.

Graaf, Maartje. M. A. de., Hindriks, Frank. A., and Hindriks, Koen. V. (2021).Who Wants to Grant Robots Rights? In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. New York, NY, USA: Association for Computing Machinery, 38–46. doi:10.1145/3434074.3446911

Gunkel, D. (2020). 2020: The Year of Robot Rights. *The MIT Press Reader* (blog). Available at: https://thereader.mitpress.mit.edu/2020-the-year-of-robot-rights/ (Accessed June 29, 2021).

Gunkel, D. J. (2018b). *Robot Rights*. Cambridge, Massachusetts: The MIT Press.

Gunkel, D. J. (2018a). The Other Question: Can and Should Robots Have Rights? *Ethics Inf. Technol.* 20 (2), 87–99. doi:10.1007/s10676-017-9442-4

The Right to Life Questioned. Introductory Remarks." (2019b). In Euthanasia, Abortion, Death Penalty and Religion - The Right to Life and Its Limitations: International Empirical Research, edited by H-G. Ziebertz and F. Zaccaria, 1–12. *Religion and Human Rights*. Cham: Springer International Publishing. doi:10.1007/978-3-319-98773-6_1

H-G. Ziebertz and F. Zaccaria (2019a). in Euthanasia, Abortion, Death Penalty And Religion - the Right To Life And its Limitations: International Empirical Research. *Religion and Human Rights* (Springer International Publishing). doi:10.1007/978-3-319-98773-6

Harris, J., and Reese Anthis, J. (2021). The Moral Consideration of Artificial Entities: A Literature Review. ArXiv:2102.04215 [Cs], January. Available at: http://arxiv.org/abs/2102.04215.

Hildt, E. (2019). Artificial Intelligence: Does Consciousness Matter? *Front. Psychol.* 10. doi:10.3389/fpsyg.2019.01535

Himma, K. E. (2009). Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent? *Ethics Inf. Technol.* 11 (1), 19–29. doi:10.1007/s10676-008-9167-5

Kant, I. (1997). "Lectures on Ethics," in *The Cambridge Edition of the Works of Immanuel Kant*. Editors P. Heath and J. B. Schneewind (Cambridge: Cambridge University Press). Translated by Peter Heath. doi:10.1017/CBO9781107049512

Kingwell, M. (2020). "Are Sentient AIs Persons?," in *The Oxford Handbook of Ethics of AI*. Editors M. D. Dubber, F. Pasquale, and S. Das, 324–342. doi:10.1093/oxfordhb/9780190067397.013.21

Kirk, Robert. (2021). "Zombies.," in *The Stanford Encyclopedia of Philosophy*. Editor E. N. Zalta (Metaphysics Research Lab, Stanford University). Available at: https://plato.stanford.edu/archives/spr2021/entries/zombies/.

Kurki, V. A. J. (2019). *A Theory of Legal Personhood. A Theory of Legal Personhood*. Oxford University Press. Available at: https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198844037.001.0001/oso-9780198844037?fbclid=IwAR3k8d4Z7s82Imk190A_xzh9pOFpuCY7N96MinwA53pImTMIeouWh4iiHS4. doi:10.1093/oso/9780198844037.001.0001

Levy, D. (2009). The Ethical Treatment of Artificially Conscious Robots. *Int. J. Soc. Robotics* 1 (3), 209–216. doi:10.1007/s12369-009-0022-6

Lewna, Andrzej. (2018). Obiektywizacja Odpowiedzialności Za Lekkomyślność W Prawie Karnym Anglii I Walii (Spojrzenie Komparatystyczne). *Czasopismo Prawa Karnego i Nauk Penalnych* 2, 67–88., no.

Lima, G., Kim, C., Ryu, S., Jeon, C., and Cha, M. (2020). Collecting the Public Perception of AI and Robot Rights. *Proc. ACM Hum.-Comput. Interact.* 4 (CSCW2), 1–24. doi:10.1145/3415206124

McArthur, N. (2017). "The Case for Sexbots.," in *Robot Sex: Social and Ethical Implications*. Editors John. Danaher and Neil. McArthur (The MIT Press). Available at: https://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262036689.001.0001/upso-9780262036689-chapter-004. doi:10.7551/mitpress/9780262036689.003.0003

Mosakas, K. (2020). On the Moral Status of Social Robots: Considering the Consciousness Criterion. *AI Soc.* doi:10.1007/s00146-020-01002-1

Naess, A. (1973). The Shallow and the Deep, Long-range Ecology Movement. A Summary*. *Inquiry* 16 (1–4), 95–100. doi:10.1080/00201747308601682

Nijssen, S. R. R., Müller, B. C. N., Baaren, R. B. v., and Paulus, M. (2019). Saving the Robot or the Human? Robots Who Feel Deserve Moral Care. *Soc. Cogn.* 37 (1), 41–S2. doi:10.1521/soco.2019.37.1.41

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Illustrated edition. London ; New York: Rowman & Littlefield Publishers.

Nyholm, S. (2018). The Ethics of Crashes with Self-Driving Cars: A Roadmap, I. *Philos. Compass* 13 (7), e12507. doi:10.1111/phc3.12507

O'Malley, J. (2019). The No Drone Zone. *Eng. Tech.* 14 (2), 34–38. doi:10.1049/et.2019.0201

Panagiotarakou, E. (2016). Who Loves Mosquitoes? Care Ethics, Theory of Obligation and Endangered Species. *J. Agric. Environ. Ethics* 29 (6), 1057–1070. doi:10.1007/s10806-016-9648-1

Pietrzykowski, T. (2018). *Personhood beyond Humanism: Animals, Chimeras, Autonomous Agents and the Law*. doi:10.1007/978-3-319-78881-4

Putman, H., and Putnam, H. (1964). Robots: Machines or Artificially Created Life? *J. Philos.* 61 (21), 668–691. doi:10.2307/2023045

Romano, D., Donati, E., Benelli, G., and Stefanini, C. (2019). A Review on Animal-Robot Interaction: from Bio-Hybrid Organisms to Mixed Societies. *Biol. Cybern* 113 (3), 201–225. doi:10.1007/s00422-018-0787-5

Ryland, H. (2021). It's Friendship, Jim, but Not as We Know it: A Degrees-Of-Friendship View of Human-Robot Friendships. *Minds & Machines*. doi:10.1007/s11023-021-09560-z

Sandini, G., and Sciutti, A. (2018). Humane Robots—From Robots with a Humanoid Body to Robots with an Anthropomorphic Mind. *ACM Trans. Human-Robot Interaction* 7, 1–7. doi:10.1145/3208954

Schröder, W. M. (2020). *Robots and Rights: Reviewing Recent Positions in Legal Philosophy and Ethics*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper ID: 3794566, Available at: https://papers.ssrn.com/abstract=3794566.

Smids, J. (2020). Danaher's Ethical Behaviourism: An Adequate Guide to Assessing the Moral Status of a Robot? *Sci. Eng. Ethics* 26 (5), 2849–2866. doi:10.1007/s11948-020-00230-4

Smith, J. K. (2021). *Robotic Persons: Our Future with Social Robots*. Westbow Press.

Sparrow, R. (2004). The Turing Triage Test. *Ethics Inf. Technol.* 6 (4), 203–213. doi:10.1007/s10676-004-6491-2

Stone, C. D. (2010). *Should Trees Have Standing?: Law, Morality, and the Environment*. Third Edition. Oxford, New York: Oxford University Press.

Turner, J. (2018). *Robot Rules: Regulating Artificial Intelligence*. Palgrave Macmillan.

Umbrello, S., and Sorgner, S. L. (2019). Nonconscious Cognitive Suffering: Considering Suffering Risks of Embodied Artificial Intelligence. *Philosophies* 4 (2), 24. doi:10.3390/philosophies4020024

Véliz, C. (2021). Moral Zombies: Why Algorithms Are Not Moral Agents. *AI Soc.* doi:10.1007/s00146-021-01189-x

Wilks, M., Caviola, L., Kahane, G., and Bloom, P. (2021). Children Prioritize Humans over Animals Less Than Adults Do. *Psychol. Sci.* 32 (1), 27–38. doi:10.1177/0956797620960398

Wróbel, W., and Zoll, A. (2017). in *Kodeks Karny. Część Szczególna. Tom II. Komentarz Do Art* (Warszawa, 117–196.

Wróbel, W., and Zoll, A. (2014). Polskie Prawo Karne: Część Ogólna. *Wyd. 3. Kraków*. Społeczny Instytut Wydawniczy Znak.

W. Wróbel, W. Zontek, and W. Adam (2014). in Kodeks Karny: Przepisy Dwujęzyczne = Criminal Code. *Stan Prawny Na 5 Listopada 2014 R. Z Uwzględnieniem Zmian Wprowadzonych Ustawą Z Dnia 27 Września 2013 R. O Zmianie Ustawy-Kodeks Postępowania Karnego Oraz Niektórych Innych Ustaw (Dz.U. Poz. 1247), Które Wejdą W Życie 1 Lipca 2015 R* (Warszawa: Lex a Wolters Kluwer business).

Zontek, W. (2018). *Modele Wyłączania Odpowiedzialności Karnej*. Kraków.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/frobt.2021.719944

On the Social-Relational Moral Standing of AI: An Empirical Study Using AI-Generated Art

Gabriel Lima [1,2], Assem Zhunis [1,2], Lev Manovich [3]* and Meeyoung Cha [1,2]*

[1]School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Korea, [2]Data Science Group, Institute for Basic Science, Daejeon, Korea, [3]The Graduate Center, City University of New York, New York, NY, United States

The moral standing of robots and artificial intelligence (AI) systems has become a widely debated topic by normative research. This discussion, however, has primarily focused on those systems developed for social functions, e.g., social robots. Given the increasing interdependence of society with nonsocial machines, examining how existing normative claims could be extended to specific disrupted sectors, such as the art industry, has become imperative. Inspired by the proposals to ground machines' moral status on social relations advanced by Gunkel and Coeckelbergh, this research presents online experiments ($\sum N = 448$) that test whether and how interacting with AI-generated art affects the perceived moral standing of its creator, i.e., the AI-generative system. Our results indicate that assessing an AI system's lack of mind could influence how people subsequently evaluate AI-generated art. We also find that the overvaluation of AI-generated images could negatively affect their creator's perceived agency. Our experiments, however, did not suggest that interacting with AI-generated art has any significant effect on the perceived moral standing of the machine. These findings reveal that social-relational approaches to AI rights could be intertwined with property-based theses of moral standing. We shed light on how empirical studies can contribute to the AI and robot rights debate by revealing the public perception of this issue.

Keywords: artificial intelligence, moral standing, moral status, agency, experience, patiency, art, rights

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Maartje M. A. De Graaf,
Utrecht University, Netherlands
Damien Williams,
Virginia Tech, United States

***Correspondence:**
Lev Manovich
manovich.lev@gmail.com
Meeyoung Cha
mcha@ibs.re.kr

Specialty section:
This article was submitted to
Ethics in Robotics and
Artificial Intelligence,
a section of the journal
Frontiers in Robotics and AI

Received: 03 June 2021
Accepted: 20 July 2021
Published: 05 August 2021

Citation:
Lima G, Zhunis A, Manovich L and
Cha M (2021) On the Social-Relational
Moral Standing of AI: An Empirical
Study Using AI-Generated Art.
Front. Robot. AI 8:719944.
doi: 10.3389/frobt.2021.719944

1 INTRODUCTION

As robots and artificial intelligence (AI) systems become widespread, scholars have questioned whether society should have any responsibility towards them. This inquiry, also called the "robot rights" debate (Gunkel, 2018b), comprehensively questions whether these systems matter morally, i.e., whether a certain level of moral standing should be granted or recognized to them. Scholars have expressed a plurality of views on this topic. Those who oppose the prospect denounce the idea by arguing that these entities are ontologically different from humans (Miller, 2015). Others argue that this debate occurs at the expense of more salient moral issues (Birhane and van Dijk, 2020) and could lead to social disruption (Bryson, 2018). In contrast, some scholars propose that robots and AI systems should matter morally if they develop consciousness or sentience (Torrance, 2008). Even if they don't become conscious, society might choose to protect AI and robots to discourage immoral human behavior (Darling, 2016).

This research is motivated by the proposals advanced by Gunkel and Coeckelbergh, both of whom advocate a social-relational perspective to the robot rights debate. Gunkel (2018a) proposes that moral status is grounded on social relations rather than an entity's ontology, such that automated systems could matter morally in the face of social interactions. In a similar vein, Coeckelbergh (2020b) argues that society could give these entities moral standing due to their extrinsic value to humans and suggests that these entities could be granted indirect moral status according to how much humans value them.

The AI and robot rights discussion has been mostly restricted to normative research. Few empirical studies have examined the public attitude towards these systems' moral standing (Lima et al., 2020; de Graaf et al., 2021). These studies have also not addressed specific perspectives advanced by previous normative work. This paper thus investigates whether social-relational approaches to this debate could be extended to a significant nonsocial robotics context, namely AI-generated art. AI-generative systems have achieved impressive results in generating a wide range of image styles (Karras et al., 2019; Goodfellow et al., 2014). Some of these images have been auctioned in the real world for remarkable prices (Cohn, 2018; Ives, 2021). Considering the social dimension of art, we inquire whether interacting with AI-generated art influences the perceived moral status of its creator, i.e., the AI-generative system.

After carefully selecting a series of AI-generated paintings (Experimental Setting, N = 45; **Section 4**), we conducted two studies inspired by the social-relational approaches advanced by Gunkel (Study 1, N = 140; **Section 5**) and Coeckelbergh (Study 2, N = 263; **Section 6**). Study 1 inquired whether interacting with AI-generated art modifies how participants perceive an AI systems' agency and patiency through a mind perception questionnaire (Gray et al., 2007). Study 2 examined whether highlighting an AI system's extrinsic value by undervaluing or overvaluing its outputs affects its perceived agency, patiency, and moral status.

Both studies show that participants deemed AI-generative systems as able to create and experience art to a significant level. Study 1 identified that nudging participants to think about an AI system's "mind" negatively influenced how they judged its artwork; this indicates that ontological considerations could play a role in interactions with non-human entities. Moreover, Study 2 found that people shown overvalued AI-generated images may undermine its creator's agency compared to other control conditions. However, none of the studies suggested that interacting with AI-generated art would influence people's perception of the AI system's moral standing. Collectively, our results reveal that considerations about the mind of non-humans could be intertwined with social-relational theses of their moral standing.

We discuss how studies like ours can contribute to the robot rights debate by obtaining empirical data supporting or challenging existing normative proposals. Scholars posit that public perceptions of AI systems could partially shape their development, use, and regulation (Cave and Dihal, 2019). Studies such as ours can thus inform future discussions on how the general public perceives AI's and robots' moral and social standing. We also propose future research directions, such as understanding how ontological considerations could play a role in human-robot interactions and whether our results extend to other environments where AI and robots are deployed.

2 BACKGROUND

2.1 Moral Status of Artificial Intelligence and Robots

Extensive literature has questioned who should be responsible for the actions of artificial intelligence (AI) and robotic systems. Some scholars propose the existence of a responsibility gap, where no entity can be appropriately held responsible for harms caused by these entities (Matthias, 2004; Asaro, 2016). Others argue that worries about a responsibility gap are overstated (Tigard, 2020) and designers should instead proactively take responsibility for their creations (Johnson, 2015). The discussion surrounding the responsibility gap (or its nonexistence) questions AI systems' moral agency, i.e., their capacity to do right or wrong. In this research, we instead follow the perspective that asks whether these systems can be subjects of rights and wrongs, i.e., whether they can (and should) be moral patients (Gunkel, 2012).

While a moral agent can act morally and possibly be deemed responsible for its actions, to be a moral patient implies that society has responsibilities towards it (Bryson, 2018). Moral patients have a certain moral status, hence suggesting that they have legitimate interests that other agents should consider, i.e., there are constraints on how one treats a moral patient (Gordon, 2020). Extensive philosophical literature has debated which conditions ground moral status. A common perspective is that moral patiency (and agency) depends on an entity satisfying specific properties (Coeckelbergh, 2014). Some scholars argue that sentience and consciousness are necessary conditions for moral patiency (Bernstein, 1998). Nevertheless, these views are rarely agreed upon, particularly in the literature discussing the moral status of non-humans (Gellers, 2020).

The debate around the moral patiency of AI systems and robots has often been framed under the umbrella of "robot rights" (Gunkel, 2018b). This setting relies on the fact that high moral status (e.g., moral patiency) grounds moral personhood, which in turn ascribes or recognizes an entity's moral rights (Gordon, 2020). The robot rights literature challenges the institutions that sort entities by type (e.g., humans, non-human animals, artifacts) and put humans on top. Scholars have argued that reinterpreting the distinction between "who" and "what" may encourage a more respectful, participatory, and dignified social order (Estrada, 2020).

Although the debate's title might suggest that scholars only propose moral status for embodied systems, research indicates that both robots and (nonphysical) AI systems could have their moral patiency recognized (e.g., see Bryson (2018); Lima et al. (2020)). Throughout this paper, we refer to "robot rights" for consistency with previous work on the topic but do not necessarily restrict our discussion to embodied systems. The series of studies covered by this research specifically address systems without any physical presence in the world, i.e., AI-generative models, and we often use "AI" and "robots" as synonyms.

Some scholars opposed to robot rights argue that its mere conception is unthinkable and should be denounced. For instance, Birhane and van Dijk (2020) argue that this debate occurs at the expense of more urgent ethical issues, such as privacy and fairness, and should be avoided at all costs. That is not to say that all scholars who oppose robots and AI systems with any moral status discard its possibility. Bryson (2018), for instance, recognizes that such systems could be accorded rights but opposes it. Bryson argues that creating systems that could be granted certain moral status is bound to conflict with a coherent ethical system and thus should be avoided.

Another series of arguments against recognizing automated agents' moral status relies on their incompatibilities with what authors defend to be moral patiency preconditions. Miller (2015) has argued against robot rights under the justification that robots are ontologically different from humans. Being created for a specific purpose, robots are not brought into the world similarly to humans. Miller defends that humans' lack of purpose lays the foundation of their rights, as they allow humans to discover their purpose. While this argument defends that granting robots and AI systems certain moral status should be denounced regardless of whether they satisfy specific properties, other scholars are disposed to granting or recognizing robots' and AI's rights if (and only if) they develop them. Torrance (2008) is one author that is open to granting moral status to automated agents if they become conscious or sentient. A distinct approach has been put forth by Danaher (2020), who proposes to use behavioral inferences as evidence of the ontological attributes that ground moral status. Such proposal posits that automated agents could be granted significant moral status if they behave similarly enough to entities with high moral status.

Various authors' perspectives to the discussion of AI and robot rights propose to ground these systems' moral patiency not on themselves but on those who interact with them. This indirect approach often suggests protecting automated agents for the sake of humans. For instance, Darling (2016) defends that society should protect social robots from cruelty to not promote such immoral behavior in human-human interactions. In a similar vein, Nyholm (2020) argues that we should respect anthropomorphized robots' apparent humanity out of respect for human beings' humanity. Friedman (2020) reinterprets the standard dyadic conception of morality and defends the protection of perceived robotic moral patients by viewing humans as both moral agents and patients of their actions towards robots. A similar approach has also been put forward by Coeckelbergh (2020a), who argues that engaging in immoral behavior towards social robots could damage an agent's moral character (i.e., its virtue), and thus should be avoided.

The present research builds upon the social-relational perspectives to robot rights put forth by Gunkel and Coeckelbergh. Inspired by the relational turn in ethics concerning non-human animals (Taylor, 2017), humans (Levinas, 1979), and the environment (Naess, 2017), both authors argue against property-based conceptions of moral patiency and defend instead that social relations ground moral status. Gunkel (2018b) argues for a direct approach to robot rights such that moral status is grounded on one's response to a social encounter with a robotic other. The author defends that moral persons are not defined by their ontological attributes but by how they engage in social relations. As Gunkel (2012) himself puts it, "moral consideration is decided and conferred not based on some pre-determined ontological criteria [. . .] but in the face of actual social relationships and interactions."

Coeckelbergh's perspective differs from Gunkel's in that it gives indirect moral standing to robots or AI systems "based on the ways humans [. . .] relate to them" (Coeckelbergh, 2020b). Although also relying on how humans interact with automated agents, his argument posits that their moral standing should instead be grounded on their extrinsic value to humans (Coeckelbergh, 2010). If humans, who are valuable per se, value robots and AI systems, the latter could also be deemed morally valuable based not on themselves but on the entity who ascribes their value. We return to these social-relational approaches to AI and robot rights in **Section 3** when motivating our series of empirical studies on people's perception of AI systems' moral standing.

2.2 Mind Perception Theory

The conceptions of moral patiency (and agency) presented above rely on philosophical interpretations of robots' and AI systems' moral standing. A different perspective has been put forward by moral psychology research, which often questions how people perceive entities' moral status under the Mind Perception Theory. Extensive research (as reviewed by (Gray et al., 2012)) has underscored the importance of people's ascription of mental capacities in moral judgments and how it maps onto attributions of moral agency and patiency.

A widely used conception of mind perception is that people perceive agents' and patients' minds in two distinct dimensions (Gray et al., 2007). The first dimension accounts for entities' capacities to feel fear, pain, be conscious, and experience other related abilities. Entities perceived to have high levels of this dimension of mind are deemed to have high *experience*, which studies suggest to correlate with the conferring of moral rights (Waytz et al., 2010). The second dimension of mind perception—termed *agency*—includes the capacity of self-control, morality, planning, thought, and others notions related to an entity's moral agency. Previous research has observed perceived agency to be linked to attributions of responsibility Gray et al. (2007).

Mind perception in the context of robots and AI systems has received significant attention in previous work. Gray et al. (2007) have found robots being ascribed moderate levels of agency and low levels of experience. In the context of economic games, Lee et al. (2021) have observed electronic agents being ascribed moral standing if systems were manipulated to possess high agency and patiency traits. Previous work has also found systems expressing emotions (e.g., with high experience) being offered larger amounts of money in economic exchanges than their low-experience counterparts (de Melo et al., 2014). In summary, previous research broadly suggests that people's ascription of agency and experience to automated agents plays a role in their interaction with these systems. Building upon the aforementioned social-relational approach to electronic agents' moral standing, we instead inquire whether interacting with these systems influences perceptions of their patiency (and agency), i.e., how people perceive their mind and corresponding moral status.

2.3 Artificial Intelligence-Generative Models

Much of the work on robots' moral status covers those systems developed for social functions, e.g., social robots. Nevertheless, we note that many of these systems, embodied or not, are not necessarily developed with sociality in mind. Robots and AI systems are currently deployed in various environments, ranging from industrial hangars to decision-making scenarios (e.g., loan and bail decisions). In this study, we distinctively investigate the social-relational approach to electronic agents' moral standing in the context of AI-generative models.

Extensive research in computer science has been devoted to developing AI-generative models. A wide range of systems have achieved impressive results in the generation of images (Goodfellow et al., 2014; Ramesh et al., 2021), text (Brown et al., 2020), music (Dhariwal et al., 2020), and even patents (Porter, 2020). AI-generated images have received considerable attention by the field, and philosophers have even questioned whether they could be considered art and have defended an open perspective to the possibility of "machine creativity" (Coeckelbergh, 2017).

The deployment of AI-generative systems has raised many ethical and legal questions. Concerned with the environmental and social costs of text-generation models, Bender et al. (2021) have urged researchers to consider the negative societal effects of large language models. AI-generative systems have also posed questions as to who should hold the copyright, intellectual property rights, and authorship of their outputs. Eshraghian (2020) has discussed how "artificial creativity" results from many actors' efforts and thus poses critical challenges to copyright law. Abbott (2020) has defended that AI systems should be considered authors of their creations so that their creativity can be legally protected. Turner (2018) has gone even further and discussed how AI systems themselves might hold the copyright of their outputs.

Image generation by AI systems has also received considerable attention from the general public. A portrait generated by an AI-generative model was sold for over $430,000 in 2018 (Cohn, 2018), raising questions about the value of "machine creativity." More recently, a self-portrait of Sophia, the robot which has been granted honorary citizenship in Saudi Arabia, was sold for nearly $700,000 under the premise of it being the first human-robot collaborative art to be auctioned (Ives, 2021). Previous research has also questioned how people perceive art-generated art. Epstein et al. (2020) have shown how people might attribute responsibility for creating realistic paintings to the AI system that generated it, particularly if it is described in an anthropomorphized manner. In a similar vein, Lima et al. (2020) found online users to only marginally denounce the idea of an AI system holding the copyright of its own generated art. Other studies found AI-generated art being evaluated unfavorably vis-a-vis their human-created counterparts (Hong and Curran, 2019; Ragot et al., 2020), even though people do not seem to be able to differentiate between them (Köbis and Mossink, 2021; Gangadharbatla, 2021).

2.4 Art as a Social Practice

The present study expands on the social-relational approaches to AI systems' moral standing in a distinctive environment that was yet to be explored by the literature: AI-generated art. While art is not social in the same way as the social robotics perspective commonly studied by scholars discussing robot rights, art production and evaluation have been often understood as a social process where many entities come together to create what one would call art.

Sociologists of culture have developed a social understanding of the arts under which the artistic production and assignment of value are viewed as social processes involving assistants, curators, galleries, museums, art critics, and many others. The artist is viewed as only one participant of this social undertaking. Many art historians and other humanities scholars also focus on the social aspects of art by showing how artistic canons evolved (i.e., what artists were recognized as "great" was changing), and how many marginalized artists (e.g., women and people of color) were excluded from the history of art (Nochlin, 1971).

One important concept developed first in sociology that later became the common-sense view of art professionals is the "art world." The art world includes everyone who participates in creating, funding, promoting, exhibiting, writing about, buying, and selling visual art. Art worlds are numerous and extensive by comprising different networks of people. What counts as "art" in each world can also be different. As discussed by Becker (2008), an art world is "the network of people whose cooperative activity, organized via their joint knowledge of conventional means of doing things, produces the kind of artworks that the art world is noted for." Both the actual objects of art and their meanings result from collective activities, shared understandings, and accepted conventions and norms.

People's perception of and interaction with art can thus be viewed as a social phenomenon. Rather than seeing our reactions to art as being completely individual and unique, we may assume that they are in part collective—e.g., people with similar backgrounds living in a particular period may have similar tastes. The influential theory in sociology of culture developed by Bourdieu (1984) indeed proposes that people's taste in the arts is related to their socioeconomic status.

This social paradigm of the arts posits that those who create, evaluate, buy, sell, and interact with art are intertwined in understanding what art is in each art world. The inclusion of AI systems into this environment raises the question of how objects of art and their meaning might be altered in the face of AI-generated art. This revolution might change what society views as art and who people regard as artists that should be included in this artistic social network. We approach this question similar to those who discuss the moral standing of AI systems. Alongside questioning who should be included in the circle of moral patients, we inquire how people embrace AI-generative systems in their art world. We thus question whether interacting with "art" generated by AI systems can influence people's attribution of moral and artistic status to generative systems.

3 SOCIAL-RELATIONAL ETHICS FOR ROBOTS AND ARTIFICIAL INTELLIGENCE

Research on the mind perception of AI has centered on how people's preconceptions of these systems' agency and patiency influence future human-machine interactions (e.g., see Lee et al. (2021); de Melo et al. (2014)). However, the social-relational approach to "robot rights" inverts this relationship and instead argues that interacting with automated agents affects how people perceive their moral status. For instance, Gunkel's proposal of social-relational ethics for grounding the moral status of robots views moral patiency as a result of social interactions, under which people are "obliged to respond [to entities] even before we know anything at all about them and their inner working" (Gunkel, 2018b). Gunkel asserts that moral status does not depend on what the other is or how it came to be but instead emerges from how we respond to "the face of the other" (Gunkel, 2018a).

Gunkel (2018b) discusses how one may anticipate an anthropocentric perspective of an entity's face by turning it "into a kind of ontological property." Instead, the author interprets this face to include other entities, such as animals, the environment, technologies, and surely robots. In this work, we expand on this idea and inquire how people respond to the "face" of an AI-generative model. These systems do not have what one would call a face one can respond to but rather output creations that can be interacted with. Study 1 covered by this research questions whether people interacting (i.e., responding) to AI-generative art (i.e., the model's "face") influences how they ascribe moral status to its creator.

Coeckelbergh (2010) similarly states that "moral significance resides neither in the object nor in the subject, but the relation between the two," suggesting that moral status can only be grounded in dynamic social relations. The author highlights that studying robots' moral considerations must account for how they are deployed and how people might interact with them. In contrast to Gunkel's, Coeckelbergh's view on how social-relational ethics can ground robots' moral status does not rely on how one might respond to electronic agents per se. It instead focuses on how others might value and interact with them, i.e., their extrinsic value.

Coeckelbergh (2020b) has proposed a set of conditions that could sufficiently ground a certain level of indirect moral standing to social robots. These conditions cover how immoral interactions with social robots could denigrate one's virtue (see also Coeckelbergh (2020a)) and how they could conflict with human-robot relationships. The present research adapts one of these conditions to a nonsocial robot environment. Coeckelbergh proposes that social robots could be given moral standing "if the human user has a (one-directional) relationship to the robot and has developed feelings of attachment and empathy towards the robot" (Coeckelbergh, 2020b). We expand on this view and inquire whether others' under- or overvaluation of an AI-generative model's outputs, i.e., whether human users have developed feelings of value towards an AI system, could ground this system's perceived moral status in Study 2.

It should be noted that the present research's case study broadens the usual setting discussed by much of the literature on automated agents' moral standing. Coeckelbergh (2020b) and Darling (2016), for instance, develop their arguments in the context of robots intentionally designed to be integrated into human social environments, i.e., social robots. As mentioned above, however, AI systems are not only deployed in social settings, and scholars have questioned whether they should be granted moral standing in diverse environments (Bryson, 2018; Turner, 2018; Lima et al., 2020). We approach this inquiry through the lens of social interpretations of art, under which artists, curators, galleries, and even laypeople contribute to creating a shared understanding of art, i.e., an art world. This research does not aim to debunk or confirm any of the social-relational approaches to robot rights; it instead seeks to provide a distinct and empirical perspective to the debate.

We present two studies aimed at understanding how social-relational approaches to robots' moral standing pertain to the context of AI-generative art. We first carefully selected a series of AI-generated images (similar to paintings produced in modern art) that online users could not discern as either human-created or AI-generated. These paintings were used in subsequent studies, and we make them available for future research. Study 1 was influenced by Gunkel's approach to "robot rights" and evaluated whether interacting with AI-generated images affects how people ascribe patiency and agency to their creator. Finally, Study 2 addressed Coeckelbergh's proposal of electronic agents' indirect moral standing by examining whether others' under- or overvaluation of AI-generated art influences an AI system's perceived moral status. All studies had been approved by the Institutional Review Board (IRB) of the first author's institution.

4 EXPERIMENTAL SETTING

Our experiments presented a series of AI-generated art-looking images to participants and explored whether interacting with AI systems' outputs influences subsequent ascription of moral status. For that, we employed a state-of-the-art model named StyleGAN2 (Karras et al., 2019) to generate images. StyleGAN2 is based on the Generative Adversarial Network (GAN) architecture (Goodfellow et al., 2014), which consists of two distinct deep neural networks, a generator and a discriminator, that compete with each other during the training process. The generator learns to output data that looks similar to the training set and aims to deceive the discriminator. In contrast, the discriminator tries to distinguish between outputs by the model and the training set's data. This model architecture has achieved impressive results in a wide range of tasks, ranging from the generation of paintings (Karras et al., 2019) and faces (Karras et al., 2017) to style transfer between images (Zhu et al., 2017).

We generated images using a pre-trained StyleGAN2 implementation available on Github (Baylies, 2020). This model had been trained on a subset of the WikiArt dataset containing over 81,000 paintings. After obtaining an initial set of 200 images, one of the authors with extensive art training selected a subset of 58 images based on their authenticity and quality. We then presented

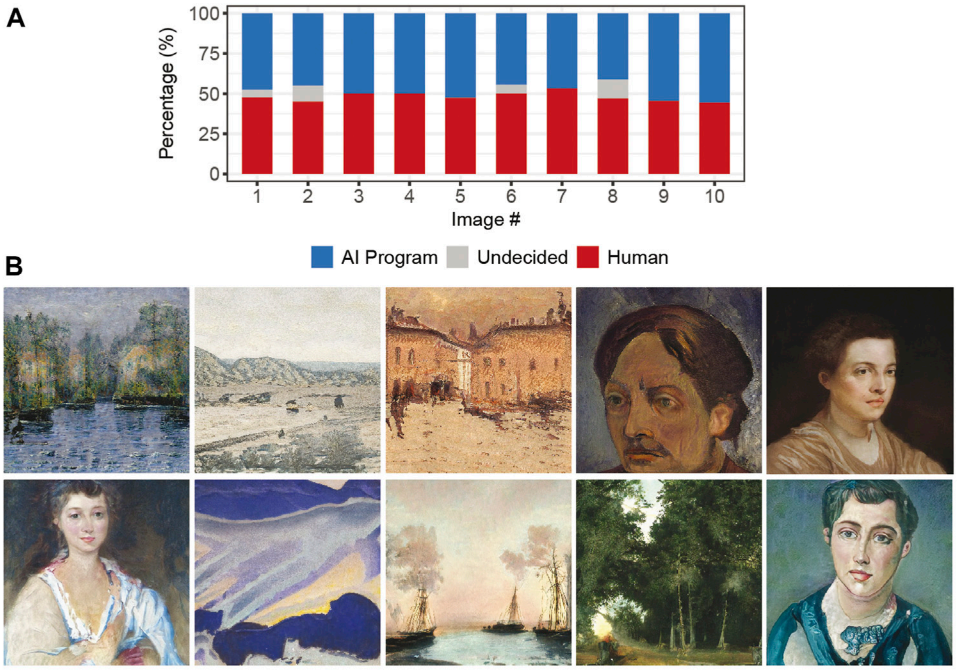

FIGURE 1 | Distribution of respondents' judgments of the top-10 images selected in our preliminary study for Studies 1 and 2 **(A)**. Selected images used in Studies 1 and 2 **(B)**.

the generated images to participants, who were asked to distinguish which images were generated by AI.

4.1 Methods

After agreeing to the research terms, study participants were told that they would be presented with a series of images generated by AI systems and human artists. Note that all images had been generated by the AI-generative model described above. Participants were instructed to indicate who they thought created each image—an AI program or a human artist. Participants were successively shown a random subset of 20 images in random order. Participants also had the option to indicate that they were unsure about its creator for each image. After evaluating all 20 images, participants were debriefed that an AI model had generated all images.

4.2 Participants

We recruited 45 respondents (22 men, 21 women, two others; 26 younger than 35 years old) through the Prolific crowdsourcing platform (https://www.prolific.co/; Palan and Schitter (2018)). Participants were required to have completed a minimum of 50 tasks in Prolific with at least a 95% approval rate. All respondents were United States nationals and were compensated $0.87 for the study.

4.3 Results

We chose images that were considered most ambiguous based on participants' ratings. This decision was made by the fact that GAN-based models are intentionally modeled to deceive a discriminator. These models' training process aims to teach a generator how to output ambiguous images that one cannot discriminate as either real (i.e., human-created) or artificial (i.e., AI-generated). Although another option would be to choose images that participants thought were human-created, we note that doing so could have made future participants suspect the images' origin. Hence, to mitigate possible deception effects, we decided to discard images that were perceived to have been created by human artists.

None of the images had a majority of respondents being unsure about its provenance. We thus used Shannon Entropy to compute image ambiguity across responses indicating that humans or AI systems created the images. We selected the top-10 images in terms of ambiguity and used them for all subsequent studies. Of the ten images, five are landscapes, four are portraits, and one is an abstraction. Qualitative analysis of all 58 images showed that more realistic images were often perceived as human-created. On the other hand, abstractions were more frequently viewed as AI-generated. **Figure 1A** presents the distribution of responses for the selected images, and **Figure 1B** shows them. All images are made available in the study's online repository for future research.

5 STUDY 1

Study 1 examined whether Gunkel's social-relational approach to electronic agents' moral standing could be applied to the context of AI-generative art. Our study employed a between-subjects design where participants interacted with AI-generated images either *before* or *after* evaluating the AI system's moral status. Our analysis controlled for previous experiences with AI-generated images and treated the difference between participants in distinct treatment groups as the effect of participants' interaction with the images in the system's perceived moral status.

5.1 Methods

After consenting to the research terms, participants were told that some AI systems are currently being used to generate images and that they would be shown a series of them created by a specific model. Each participant was randomly assigned to one of two conditions. Participants assigned to the *pre* condition first responded to a series of questions compiled from previous work on mind perception theory (Gray et al., 2007; Bigman and Gray, 2018). Participants rated an AI system that can generate images concerning their perceived agency (e.g., to what extent the AI system "is intelligent;" six questions in total, see **Supplementary Table S1**) and experience (e.g., "can experience happiness;" six questions in total). We additionally asked participants to evaluate the system's ability to create art (hereafter art agency) and experience art (hereafter art experience). All judgments were made on a 5-point scale from 0 (Not at all) to 4 (Extremely). Afterward, study participants were presented to all ten images selected in our Experimental Setting in random order. Participants were asked to evaluate each of the paintings in the range between $0 and $10,000.

Study participants assigned to the *post* condition responded to the same set of questions and art evaluations; however, in the opposite order, i.e., they first evaluated all ten images and then attended to the mind perception questionnaire. Participants did not differ in how long they spent evaluating the images (t (129.2) = 0.713, $p = 0.48$, $d = 0.12$) and rating the AI systems' moral status (t (120.2) = −1.351, $p = 0.18$, $d = 0.23$) across conditions. All participants answered a series of demographic questions at the end of the study, including whether they had received any training in computer science or art-related subjects. We also gathered responses to a modified questionnaire of NARS (Negative Attitude towards Robot Scale) (Syrdal et al., 2009), with a modified text that covered "artificial intelligence programs" instead of "robots."

5.2 Participants

Power analysis indicated that 128 participants were required for detecting a medium effect size ($\eta^2 = 0.06$) with the power of 0.80 and $\alpha = 0.05$ (Campbell and Thompson, 2012). Hence, we recruited 160 respondents through the Prolific crowdsourcing platform. After removing respondents that failed an instructed response attention check question and those who had previously participated in a study where they had to evaluate AI-generated art (i.e., had interacted with AI-generated images before), our sample consisted of 140 participants (60 women, 77 men, three others) aged between 19 and 77 years old (mean = 31.96, SD = 11.96). We enforced the same recruitment conditions and payment as in Study 1.

5.3 Results

A principal component analysis (PCA) of participants' attribution of moral status revealed two dimensions with eigenvalues larger than one (see **Supplementary Tables S1** and **S2**). After varimax rotation, the first component (termed "experience") accounted for all experience-related questions from the mind perception questionnaire with loadings greater than 0.78. The second factor (termed "agency") included all agency-related questions with loadings greater than 0.65. We thus calculated a mean attribution of experience (Cronbach's α = 0.93) and agency (α = 0.83) to the AI-generative system for each participant. Neither of the two principal components significantly accounted for art agency and experience (i.e., loadings were smaller than 0.6). These two variables were also not strongly correlated ($r = 0.404$, $p < 0.001$); we thus consider these two questions as distinct variables in our analysis.

The participants attributed moderate levels of agency (M = 1.85, SD = 0.96) and art agency (M = 2.59, SD = 1.16) to the AI-generative system. On the other hand, AI systems were rated as slightly able to experience art (M = 1.10, SD = 1.2) and were attributed almost no experience (M = 0.34, SD = 0.67). To what extent the study participants attributed agency (M_{pre} = 1.97, M_{post} = 1.74, t (136.5) = −1.427, $p = 0.15$, $d = 0.24$) and patiency (M_{pre} = 0.25, M_{post} = 0.42, t (132.9) = 1.531, $p = 0.13$, $d = 0.25$) to the AI system did not differ significantly across treatment conditions. Nevertheless, the participants attributed marginally higher levels of art agency (M_{pre} = 2.38, M_{post} = 2.77, t (125.6) = 1.981, $p = 0.05$, $d = 0.34$) and art experience (M_{pre} = 0.88, M_{post} = 1.29, t (135.1) = 2.119, $p = 0.04$, $d = 0.35$) to the generative model had they rated the system's moral status after interacting with the images.

The observations above raise the question of whether moral patiency and agency attribution differs across participants with distinct perceptions of AI-generated art, i.e., how each participant individually valued the presented images. We hence conducted an analysis of variance (ANOVA) accounting for the interaction between the study condition and the average value assigned to all images by each participant. We did not observe any significant effect of the treatment condition and its interaction with art evaluation across all dependent variables ($p > 0.05$ for all F-tests). We found the same results when controlling for respondents' attitudes towards AI and their previous knowledge of computer science and art-related subjects. We present the estimated marginal means of all dependent variables and their corresponding 95% confidence intervals in **Figure 2A**.

An exploratory analysis of how participants evaluated the set of AI-generated paintings showed a large difference between respondents in distinct groups; those evaluating the images before attending to the mind perception questionnaire perceived the images to be more valuable (M_{pre} = 2,149, M_{post} = 3,244, t (137.9) = 3.244 $p = 0.001$, $d = 0.55$). A mixed-effects model regressing participants' evaluation of all AI-generated paintings with treatment condition and image number as fixed effects indicated that respondents differed across conditions (F (1, 138) = 10.352, $p = 0.002$). We estimated marginal means across all ten images and found participants who evaluated all paintings before attending to the mind perception questionnaire to value them more highly (95% CI, M_{pre} = [1,657, 2,642], M_{post} = [2,786, 3,703], $p = 0.002$; see **Figure 2B**). We observed qualitatively

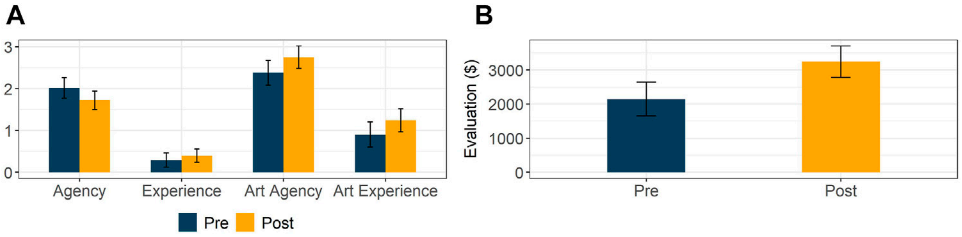

FIGURE 2 | Attribution of agency, experience, art agency, and art experience to an AI system before and after being exposed to AI-generated paintings in Study 1 **(A)**. Marginal mean evaluation across all ten images depending on treatment group **(B)**.

similar results when accounting for respondents' attitudes towards AI and their previous knowledge of computer science and art-related subjects.

5.4 Discussion

Whether participants interacted with AI-generated images before or after attributing moral agency and patiency to the system did not influence its perceived moral standing. We observed a significant difference in participants' perception of the AI system's capacity to create and experience art depending on the treatment condition. This effect, however, disappeared once we controlled for participants' attitudes towards the AI systems' outputs, i.e., the average price assigned to AI-generated art. It may well be the case that our proposed interaction with AI-generated art is not as strong a stimuli as the significant social interactions that authors defend to be crucial components of moral standing.

Nevertheless, study participants ascribed the ability to create art to the AI system although it was not described as an "artist," nor their outputs were introduced as "art." This specific artistic notion of the agency was perceived as more significant to the AI-generative system than the more general conception of agency captured by the mind perception questionnaire. In a similar vein, our results indicate that AI systems were attributed some ability to experience art even though they were not perceived to have the experience dimension of mind.

Finally, we observed a significant difference across treatment groups by expanding our analysis to how participants responded to AI-generated paintings. Even after controlling for individual variations through a mixed-effects model, AI-generated images were valued lower by participants who attributed moral standing to the AI system before interacting with its images. This result suggests that nudging participants to think about an AI system's mind (e.g., its agency and patiency) could negatively influence how much they value its outputs. That is, the act of evaluating an AI system's moral status could influence how people interact with them.

6 STUDY 2

Study 2 inquired whether Coeckelbergh's socio-relational approach to electronic agents' indirect moral status could be extended to the context of AI-generative art. The author suggests that electronic agents could be granted moral standing if others have a valuable relationship with them, i.e., one should respect these systems' interests due to their extrinsic value. Hence, our study was designed to randomly assign participants to treatment groups that show how others perceived AI-generated images, e.g., by under- or overvaluing them.

6.1 Methods

After agreeing to the research terms, participants were told that some existing AI systems could generate images and that they would be shown some examples throughout the study. Each participant was randomly assigned to one of four treatment groups. Those assigned to the *pre* condition took part in a study similar to the *pre* condition in Study 1, i.e., they attributed moral status before interacting with a series of AI-generated images. Participants allocated to the *undervalue*, *median*, and *overvalue* conditions were presented a study design similar to Study 1's *post* condition, where participants first evaluated a set of AI-generated paintings and then answered questions concerning their creator's moral status.

Study 2 differed from the previous study in that participants were shown additional information during the art evaluation step. After evaluating each of the images, participants were shown how other respondents evaluated the same painting depending on the treatment condition they were assigned to. They were subsequently asked to modify their initial evaluation if they desired to do so. Participants assigned to *pre* and *median* conditions were shown median values calculated from Study 1's responses.[1] Those in the *undervalue* and *overvalue* groups were presented to evaluations three times lower or larger than those presented in the other two conditions. This design choice aimed to elucidate the AI system's extrinsic value, which Coeckelbergh argues to be crucial for electronic agents' moral standing.

All participants responded to the same mind perception questionnaire and art-related questions from Study 1. We additionally asked participants to rate the AI-generative system's moral standing concerning six statements. Respondents

[1] Due to a programming error, median values were calculated with respect to the order images were shown to participants in Study 1. For instance, image #1's median value was determined by the median evaluation of the first image shown to each participant. Note that the image order was randomized between participants. Our study conditions should not be affected by this error, i.e., all images were overvalued or undervalued on their respective treatment conditions.

were asked to what extent the system "has legitimate interests," "can have rights," "has inherent value," "is more than just a tool," "deserves protection," and "deserves moral consideration." These questions were created after an extensive review of the recent literature addressing the moral standing of electronic agents (Gunkel, 2018a; Coeckelbergh, 2020b; Gordon, 2020). All judgments were made on a 5-point scale from 0 (Not at all) to 4 (Extremely). Participants did not differ in how long they spent evaluating the images (all $p > 0.05$ after Bonferroni corrections) and rating the AI systems' moral status (all $p > 0.05$) across conditions. Finally, participants were asked the same demographic and personal experience questions from Study 1 before completing the study.

6.2 Participants

Considering the power analysis conducted for Study 1, we decided to double the number of participants recruited for this study to account for doubled treatment conditions. We thus recruited 315 respondents through Prolific. After removing respondents that failed an attention check question similar to Study 1's and those who had previously participated in a study where they had to evaluate AI-generated art, our sample consisted of 263 participants (126 women, 134 men, three others) aged between 19 and 75 years old (mean = 34.40, SD = 12.73). Recruitment requirements and conditions were the same as in previous studies.

6.3 Results

We identified four principal components with eigenvalues larger than one by analyzing participants' ratings of the AI system's moral status (see **Supplementary Tables S3** and **S4**). The first two components accounted for all of the experience- and agency-related questions with loadings greater than 0.84 and 0.69, respectively. In a similar manner to Study 1, we calculated mean attributions of experience ($\alpha = 0.96$) and agency ($\alpha = 0.88$) for each participant. The third factor identified by the principal component analysis included five out of the six novel moral standing-related questions (with loadings greater than 0.61). In contrast, the last factor accounted for this extra item ("has inherent value," loading equal to 0.69) and art agency (loading equal to 0.87). We again kept art agency and experience as independent variables due to their low correlation ($r = 0.411$, $p < 0.0001$). We finally calculated participants' mean attribution of moral status by averaging all items proposed by this study ($\alpha = 0.86$). All results discussed below are qualitatively similar to those controlling for participants' attitudes towards AI and their previous knowledge of computer science and art-related subjects.

As a manipulation check, we analyzed whether treatment groups differed in how much participants modified their initial evaluation after seeing others' judgments. We ran a mixed-effects model regressing evaluation-change with the study condition and the image number as fixed effects. Participants' initial evaluation was included as a covariate. The results suggest that the condition to which participants were assigned played a role in how much they changed their initial evaluation (F (3, 220) = 26.684, $p < 0.001$). Pairwise comparisons between marginal means across all images show that participants presented to overvalued AI-generated art increased their initial evaluation after treatment. In contrast, those assigned to all others conditions decreased their evaluation—we note that evaluation-change did not significantly differ between the *pre*, *median*, and *undervalue* conditions (see **Figure 3A**).

Similarly to Study 1, participants attributed moderate levels of agency ($M = 1.45$, $SD = 1.00$) and art agency ($M = 2.54$, $SD = 1.27$) to the AI system, while it was rated as slightly capable of experiencing art ($M = 1.09$, $SD = 1.27$). Participants attributed low levels of experience ($M = 0.28$, $SD = 0.63$) and moral status ($M = 1.17$, $SD = 0.90$) to the automated system. Pairwise t-tests between study conditions only suggested a significant difference in the attribution of agency. After Bonferroni corrections, we observed that participants presented overvalued AI art attributed lower levels of agency to their creator than those who evaluated it before interacting with the AI-generated images ($M_{pre} = 1.86$, $M_{overvalue} = 1.31$, t (1,102) = −3.02, $p = 0.02$, $d = 0.55$; all others $p > 0.05$).

Having found non-significant differences in evaluation-change across treatments, we analyzed ANOVA models with study conditions and their interaction with the extent to which participants changed their initial evaluation (i.e., the treatment effect) as fixed effects. Respondents' average initial art evaluation was included as a covariate. There were significant differences across treatment groups for the AI system's perceived agency (F (3, 254) = 3.985. $p < 0.01$). The estimated marginal means showed higher attributions of agency by participants in the *pre* condition vis-as-vis those in the *overvalue* treatment group (95% CI, M_{pre} = [1.59, 2.13], $M_{overvalue}$ = [0.97, 1.51], $p = 0.01$; see **Figure 3B**). To what extent participants attributed all other variables did not differ across conditions ($p > 0.05$ for all F-tests).

Finally, we analyzed how differently participants evaluated the AI-generated paintings they were shown depending on the study condition they were assigned to. We ran a mixed-effects model with the experimental condition and image number as fixed effects and evaluation-change as a covariate. We included the interaction term between the study condition and the evaluation change to account for the non-significant contrasts between some treatment conditions. Here, the condition played a significant role in how participants evaluated the AI-generated images (F (3, 259) = 20.235, $p < 0.001$). As expected from the treatment condition, participants assigned to the *overvalue* condition evaluated AI-generated images more highly in comparison to those in all other conditions (95% CI, M_{pre} = [1,586, 2,375], $M_{undervalue}$ = [953, 1,667], M_{median} = [1,593, 2,271], $M_{overvalue}$ = [2,852, 3,547], all $p < 0.001$,; see **Figure 3C**). All other contrasts were not significant ($p > 0.05$).

6.4 Discussion

Similarly to Study 1, participants attributed higher levels of art-related agency and experience than their more general (and moral) counterparts to the AI-generative system. The result was again observed without explicitly introducing the AI system as an "artist" or its outputs as "art." Our results reveal that participants attributed experience, moral status, art agency, and art experience regardless of our study's nudges concerning the AI-generative model's extrinsic value. In contrast, participants showed a distinction concerning the AI system's perceived agency—overvaluing the system's outputs led to a lower perceived agency in comparison to ratings prior to interacting with AI-generated art.

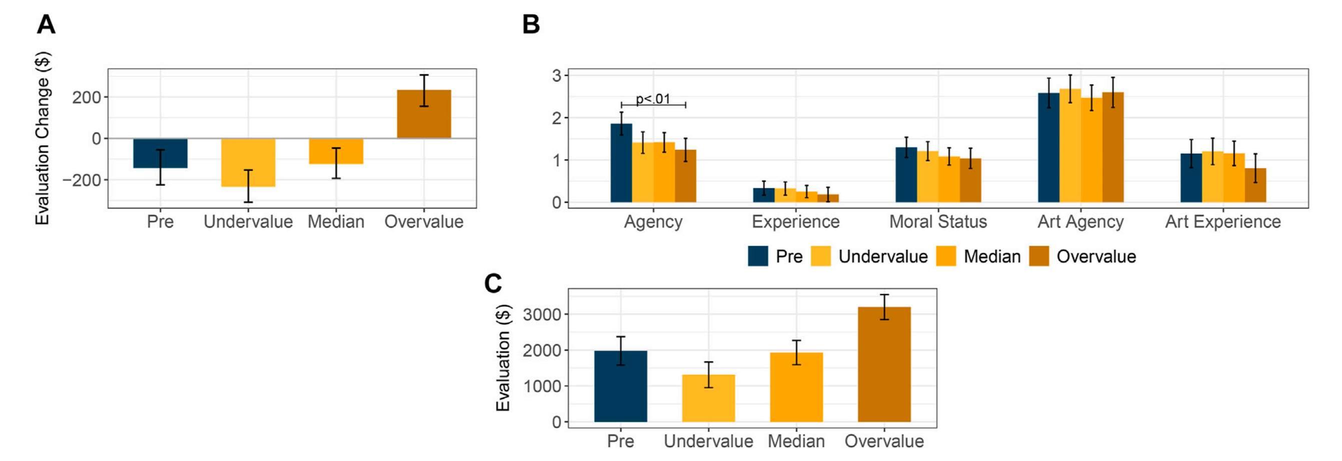

FIGURE 3 | To what extent participants modified their initial art evaluation after treatment in Study 2 **(A)**. Attributions of agency, experience, moral status, art agency, and art experience to the AI system depending on the condition participants were assigned to in Study 2 **(B)**. Marginal mean evaluation across all ten images depending on treatment group **(C)**.

We expanded Study 2 to include a novel measure of perceived moral standing independent of an entity's perceived experience covered by the mind perception questionnaire. This was done because the social-relational approach to electronic agents' moral standing challenges perspectives that defend experience-related capacities as preconditions for moral status. Nevertheless, we did not find any significant difference between treatment conditions in both attributions of experience and our proposed moral standing measure. These results corroborate our findings from Study 1 by showing that interacting with AI-generated outputs should not influence people's ascription of moral standing.

Nudging people to think about the mind of an AI system did not necessarily influence how they valued AI-generated art in Study 2. Our results instead suggest that overvaluing AI-generated art could influence how people perceive it. We hypothesize that the treatment conditions' social influence mitigated any possible effect of considerations about an AI system's mind similar to those found in Study 1. Similar to how past auctions of AI-generated art were presented to the public (Cohn, 2018; Ives, 2021), overvaluing these outputs could influence how much people value them.

7 GENERAL DISCUSSION

Inspired by Gunkel's and Coeckelbergh's social-relational approaches to robots' moral standing, we conducted two studies to understand whether a similar perspective would influence people's ascription of moral status to a nonsocial automated agent, namely an AI-generative system. We first identified a set of ten AI-generated images that were used in subsequent studies. Study 1 inquired whether interacting with these images would influence people's ascription of moral agency and patiency to their creator—as suggested by Gunkel (2018b). Study 2 asked whether highlighting an AI system's extrinsic value by undervaluing or overvaluing its images affected participants' attribution of agency, experience, and moral status, as proposed by Coeckelbergh (2020b). The current research took a novel experimental approach to the normative debate of robot rights in the context of AI-generated art.

We employed a series of measures to quantify AI systems' perceived moral (and artistic) standing. Interacting with AI-generated art did not significantly impact how participants perceived the system's ability to create art, experience art, and the experience dimension of mind in both Studies 1 and 2. The latter was measured by a mind perception questionnaire, whose measure has been shown to correlate with the recognition of moral rights (Waytz et al., 2010; Gray et al., 2007). Study 2 also showed that interacting with AI-generated art did not influence the AI system's perceived moral standing in a novel measure of moral consideration independent of the system's experience.

Study 2's participants attributed lower levels of agency to AI systems after interacting with overvalued AI-generated art. This finding suggests that seeing others overvaluing AI systems' abilities could negatively influence their perceived agency. This finding may be contrary to what one would expect. Similar to Coeckelbergh's approach to AI systems' patiency, highlighting the system's creative value by overvaluing its generated images should, at first thought, increase their perceived (artistic) agency.

Finally, Study 1 suggests that nudging participants to think about an AI systems' mind could lead to a lower appreciation of AI-generated art. A possible interpretation is that machine creativity is not valued to the same extent as its human counterparts, particularly when AI systems' lack of humanness and mind becomes apparent. As argued by some scholars, AI-generated art may lack the meaning necessary to be considered art—such meaning can only emerge from human artistic communication (Elgammal, 2020). Another possible explanation is that art is also evaluated by the effort put into its creation. More realistic images in our Experimental Setting were often attributed to human artists, while abstractions were usually viewed as AI-generated. Participants might have judged the generation process of an AI-generated art not as labor and particularly mind intensive as human-created art. As one participant has put it in an open-ended comment to our study, "knowing that an AI made it devalues [the image]."

7.1 Limitations and Future Work

Both studies have found AI-generative systems being perceived as an agent and patient to a higher level for their particular artistic abilities. Under the social paradigm of art described above, participants included AI systems in their art world. Most AI systems are proficient in a narrow task, such as generating images, and our results suggest that participants rate their agency and patiency similarly. This observation raises the question of how participants would ascribe moral status to an AI system that is explicitly described as a moral agent or patient. For instance, scholars have proposed the creation of "artificial moral agents" capable of identifying and resolving moral dilemmas (Wallach and Allen, 2008). Past research has also explored how people interact with robots described as emotional (de Melo et al., 2014; Lee et al., 2021). A future line of research could inquire how social interactions with AI systems with different abilities would affect their perceived moral standing.

Presenting participants with others' judgments of an AI system's outputs, as done in Study 2, seems to influence their evaluation negatively. Although this effect was countered by others' overvaluation of AI-generative art, which led participants to increase their initial evaluation, respondents appear to decrease their initial evaluation even if presented with other participants' median judgments. As shown by Study 1, making participants think about the AI-generative system's lack of mind decreased how much they value its outputs. Similarly, forcing participants to think more about AI-generated art influenced how much they value it. Future work may study how nudging people to think (harder) about an AI system's (lack of) mind and its outputs may influence how participants evaluate its creations.

The current research examined a growing research area, namely AI-generative models. Extensive research has been devoted to developing and improving generative systems (e.g., Ramesh et al. (2021); Brown et al. (2020)), and many of them are already deployed in the wild (Warren, 2020; Dorrier, 2021). Our results, however, may not extend to other applications of AI systems. For instance, in the context of social robots, Darling (2016) has presented a series of anecdotes suggesting that people desire to protect social robots after interacting with them. Future research in a wide range of applications is needed to explore how people might perceive AI systems' and robots' moral standing in different environments.

We have explored Gunkel's and Coeckelbergh's social-relational perspective on robots' moral standing in the context of AI-generated art. This setting was chosen for its prominence in the AI research agenda, its legal and moral issues (e.g., concerning copyright law), and the widespread attention to AI-generated art auctions worldwide. Although art does contain a social dimension, our studies' stimuli may not have simulated the social interactions proposed by both authors in their theses. Nevertheless, we empirically explored both perspectives in a setting that was yet to be comprehensively investigated by previous experimental and normative research.

Our results confront the thesis that property-based grounds for moral patiency can be entirely substituted by social-relational perspectives (Coeckelbergh, 2010) in that considerations about the mind of non-humans, i.e., a form of ontological consideration, may influence future interactions. This finding suggests that even if social-relational approaches can ground the moral standing of machines, they may not be entirely detached from the property-based views they challenge. Instead, the property and relational approaches can be intertwined in justifying moral standing, as discussed by Gellers (2020).

Our findings contribute extensively to the discussion concerning AI systems' and robots' moral status. Our results provide scholars with empirical evidence and methods that can influence future normative discussion on the topic. For instance, we found that nudging participants to think about AI systems' (lack of) mind could influence future social interactions in the context of AI-generated art, which is an important addition to the social-relational perspectives studied in this paper. We call for future research that empirically examines normative debates on AI systems' and robots' moral agency and patiency so that subsequent discussions concerning how automated agents should be included in our moral and social spheres can make fruitful progress.

DATA AVAILABILITY STATEMENT

The datasets and scripts used for analysis presented in this study can be found at https://github.com/thegcamilo/AIArt_MoralStanding.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board at KAIST. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors designed the research. GL and AZ conducted the research. GL analyzed the data. GL wrote the paper, with critical edits from LM and MC.

FUNDING

This research was supported by the Institute for Basic Science (IBS-R029-C2).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.719944/full#supplementary-material

REFERENCES

Abbott, R. (2020). *The reasonable robot: artificial intelligence and the law*. Cambridge, United Kingdom: Cambridge University Press.

Asaro, P. M. (2016). "The Liability problem for autonomous artificial agents," in 2016 AAAI Spring symposium series, Stanford, CA, 190–194.

Baylies, P. (2020). *Adapted stylegan2 github repository*. https://tinyurl.com/v5cczeun (Accessed Mar 30, 2021).

Becker, H. S. (2008). *Art worlds: updated and expanded*. California: University of California Press.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Canada: Association for Computing Machinery. 610–623.

Bernstein, M. H. (1998). *On moral considerability: An essay on who morally matters*. Oxford, United Kingdom: Oxford University Press.

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

Birhane, A., and van Dijk, J. (2020). Robot rights? let's talk about human welfare instead. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY: Association for Computing Machinery. 207–213.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Massachusetts: Harvard University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). *Language models are few-shot learners*. *arXiv preprint arXiv:2005.14165*.

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics Inf. Technol.* 20, 15–26. doi:10.1007/s10676-018-9448-6

Campbell, J. I. D., and Thompson, V. A. (2012). Morepower 6.0 for anova with relational confidence intervals and bayesian analysis. *Behav. Res.* 44, 1255–1265. doi:10.3758/s13428-012-0186-0

Cave, S., and Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nat. Mach Intell.* 1, 74–78. doi:10.1038/s42256-019-0020-9

Coeckelbergh, M. (2017). Can machines create art?. *Philos. Technol.* 30, 285–303. doi:10.1007/s13347-016-0231-5

Coeckelbergh, M. (2020a). How to use virtue ethics for thinking about the moral standing of social robots: A relational interpretation in terms of practices, habits, and performance. *Int. J. Soc. Robotics*, 1–10. doi:10.1007/s12369-020-00707-z

Coeckelbergh, M. (2010). Robot rights? towards a social-relational justification of moral consideration. *Ethics Inf. Technol.* 12, 209–221. doi:10.1007/s10676-010-9235-5

Coeckelbergh, M. (2020b). Should we treat teddy bear 2.0 as a kantian dog? four arguments for the indirect moral standing of personal social robots, with implications for thinking about animals and humans. *Minds and Machines*, 1–24.

Coeckelbergh, M. (2014). The moral standing of machines: Towards a relational and non-cartesian moral hermeneutics. *Philos. Technol.* 27, 61–77. doi:10.1007/s13347-013-0133-8

Cohn, G. (2018). *Ai art at christie's sells for $432,500*. https://tinyurl.com/yynncj53 (Accessed Mar 29, 2021).

Danaher, J. (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Sci. Eng. Ethics* 26, 2023–2049. doi:10.1007/s11948-019-00119-x

Darling, K. (2016). "Extending Legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects," in *Robot Law* (Cheltenham, United Kingdom: Edward Elgar Publishing).

de Graaf, M. M., Hindriks, F. A., and Hindriks, K. V. (2021). "Who wants to grant robots rights?," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY: Association for Computing Machinery, 38–46.

de Melo, C., Gratch, J., and Carnevale, P. (2014). *The importance of cognition and affect for artificially intelligent decision makers*, Proceedings of the AAAI Conference on Artificial Intelligence. Quebec, Canada: AAAI Press. 28.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., and Sutskever, I. (2020). *Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341.*

Dorrier, J. (2021). Openai's gpt-3 algorithm is now producing billions of words a day. https://tinyurl.com/jkc4r57u (Accessed Apr 5, 2021).

Elgammal, A. (2020). *The robot artists aren't coming*. https://tinyurl.com/z9xu54ey (Accessed Apr 9, 2021).

Epstein, Z., Levine, S., Rand, D. G., and Rahwan, I. (2020). Who gets credit for ai-generated art?. *Iscience* 23, 101515. doi:10.1016/j.isci.2020.101515

Eshraghian, J. K. (2020). Human ownership of artificial creativity. *Nat. Mach Intell.* 2, 157–160. doi:10.1038/s42256-020-0161-x

Estrada, D. (2020). Human supremacy as posthuman risk. *The J. Sociotechnical Critique* 1, 5.

Friedman, C. (2020). *Human-robot moral relations: Human interactants as moral patients of their own agential moral actions towards robots*. Berlin, Germany: Springer, Southern African Conference for Artificial Intelligence Research. 3–20. doi:10.1007/978-3-030-66151-9_1

Gangadharbatla, H. (2021). The role of ai attribution knowledge in the evaluation of artwork. *Empirical Stud. Arts*, 0276237421994697. doi:10.1177/0276237421994697

Gellers, J. C. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law (Edition 1)*. New York, NY: Routledge.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in neural information processing systems, 27*. New York, NY: Curran Associates, Inc.

Gordon, J.-S. (2020). *Artificial moral and legal personhood*. Berlin, Germany: AI & SOCIETY, 1–15.

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of mind perception. *science* 315, 619. doi:10.1126/science.1134475

Gray, K., Young, L., and Waytz, A. (2012). Mind perception is the essence of morality. *Psychol. Inq.* 23, 101–124. doi:10.1080/1047840x.2012.651387

Gunkel, D. J. (2018b). *Robot rights*. Massachusetts: MIT Press.

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. Massachusetts: MIT Press.

Gunkel, D. J. (2018a). The other question: Can and should robots have rights?. *Ethics Inf. Technol.* 20, 87–99. doi:10.1007/s10676-017-9442-4

Hong, J.-W., and Curran, N. M. (2019). Artificial Intelligence, Artists, and Art. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1–16. doi:10.1145/3326337

Ives, M. (2021). The Latest artist selling nfts? it's a robot. https://tinyurl.com/37v5ayvh (Accessed Mar 29, 2021).

Johnson, D. G. (2015). Technology with no human responsibility?. *J. Bus Ethics* 127, 707–715. doi:10.1007/s10551-014-2180-1

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). *Progressive growing of gans for improved quality, stability, and variation*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401–4410.

Karras, T., Laine, S., and Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4401–4410. (CA, United States).

Köbis, N., and Mossink, L. D. (2021). Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Comput. Hum. Behav.* 114, 106553. doi:10.1016/j.chb.2020.106553

Lee, M., Lucas, G., and Gratch, J. (2021). Comparing mind perception in strategic exchanges: Human-agent negotiation, dictator and ultimatum games. *J. Multimodal User Inter.*, 1–14. doi:10.1007/s12193-020-00356-6

Levinas, E. (1979). *Totality and infinity: An essay on exteriority*, 1. Berlin, Germany: Springer Science & Business Media.

Lima, G., Kim, C., Ryu, S., Jeon, C., and Cha, M. (2020). Collecting the public perception of ai and robot rights. *Proc. ACM Hum.-Comput. Interact.* 4, 1–24. doi:10.1145/3415206

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of Learning automata. *Ethics Inf. Technol.* 6, 175–183. doi:10.1007/s10676-004-3422-1

Miller, L. F. (2015). Granting automata human rights: Challenge to a basis of full-rights privilege. *Hum. Rights Rev.* 16, 369–391. doi:10.1007/s12142-015-0387-x

Naess, A. (2017). "The Shallow and the Deep, Long-Range Ecology Movement. A Summary *," in *The Ethics of the Environment* (England, UK: Routledge), 115–120. doi:10.4324/9781315239897-8

Nochlin, L. (1971). Why have there been no great women artists?. *feminism Vis. Cult. Read.*, 229–233.

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Maryland: Rowman & Littlefield Publishers.

Palan, S., and Schitter, C. (2018). Prolific.ac-A subject pool for online experiments. *J. Behav. Exp. Finance* 17, 22–27. doi:10.1016/j.jbef.2017.12.004

Porter, J. (2020). Us patent office rules that artificial intelligence cannot be a Legal inventor. https://tinyurl.com/2v7khzz2 (Accessed Mar 29, 2021).

Ragot, M., Martin, N., and Cojean, S. (2020). Ai-generated vs. human artworks. a perception bias towards artificial intelligence? Extended abstracts of the 2020 CHI conference on human factors in computing systems. New York, NY: Association for Computing Machinery, 1–10.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., et al. (2021). Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*

Syrdal, D. S., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2009). The negative attitudes towards robots scale and reactions to robot behaviour in a Live human-robot interaction study. Adaptive and Emergent Behaviour and Complex Systems : Procs of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2009. SSAISB, 109–115.

Taylor, S. (2017). *Beasts of burden: Animal and disability liberation*. New York: The New Press.

Tigard, D. W. (2020). There is no techno-responsibility gap. *Philos. Tech.*, 1–19. doi:10.1007/s13347-020-00414-7

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI Soc.* 22, 495–521. doi:10.1007/s00146-007-0091-8

Turner, J. (2018). *Robot rules: Regulating artificial intelligence*. Berlin, Germany: Springer.

Wallach, W., and Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. United Kingdom: Oxford University Press.

Warren, T. (2020). Microsoft Lays off journalists to replace them with ai. https://tinyurl.com/d386phkc (Accessed Apr 5, 2021).

Waytz, A., Gray, K., Epley, N., and Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends cognitive sciences* 14, 383–388. doi:10.1016/j.tics.2010.05.006

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). *Unpaired image-to-image translation using cycle-consistent adversarial networks*, 2223–2232. Proceedings of the IEEE international conference on computer vision. doi:10.1109/iccv.2017.244 CrossRef Full Text

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Robotics and AI

HYPOTHESIS AND THEORY
published: 14 September 2021
doi: 10.3389/frobt.2021.744426

Challenging the Neo-Anthropocentric Relational Approach to Robot Rights

*Henrik Skaug Sætra**

Faculty of Computer Sciences, Engineering and Economics, Østfold University College, Halden, Norway

When will it make sense to consider robots candidates for moral standing? Major disagreements exist between those who find that question important and those who do not, and also between those united in their willingness to pursue the question. I narrow in on the approach to robot rights called relationalism, and ask: if we provide robots moral standing based on how humans relate to them, are we moving past human chauvinism, or are we merely putting a new dress on it? The background for the article is the clash between those who argue that robot rights are possible and those who see a fight for robot rights as ludicrous, unthinkable, or just outright harmful and disruptive for humans. The latter group are by some branded human chauvinists and anthropocentric, and they are criticized and portrayed as backward, unjust, and ignorant of history. Relationalism, in contrast, purportedly opens the door for considering robot rights and moving past anthropocentrism. However, I argue that relationalism is, quite to the contrary, a form of neo-anthropocentrism that recenters human beings and their unique ontological properties, perceptions, and values. I do so by raising three objections: 1) relationalism centers human values and perspectives, 2) it is indirectly a type of properties-based approach, and 3) edge cases reveal potentially absurd implications in practice.

Keywords: anthropocentrism, ethics, moral standing, robots, rights, social robots, robot rights, neo-anthropocentrism

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Mark Coeckelbergh,
University of Vienna, Austria
Anne Gerdes,
University of Southern Denmark,
Denmark

Correspondence:
Henrik Skaug Sætra
henrik.satra@hiof.no

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

***Received:** 20 July 2021*
***Accepted:** 01 September 2021*
***Published:** 14 September 2021*

Citation:
Sætra HS (2021) Challenging the Neo-Anthropocentric Relational Approach to Robot Rights.
Front. Robot. AI 8:744426.
doi: 10.3389/frobt.2021.744426

INTRODUCTION

If we provide robots moral standing because of how humans relate to them, are we moving past human chauvinism, or are we merely putting a new dress on it? Questions related to moral standing go back a long way (Sætra, 2019), and they always trigger strong emotions and require that we deal with both difficult and fundamental questions. Different types of humans–demarcated by color, sex, and a range of other arbitrary attributes of questionable moral relevance–have fought tough battles for being recognized as of equal, or at least some, value. Other entities, such as animals, cannot fight for their own rights, but humans have still taken it upon themselves to fight for their rights (Regan, 2004). Even rivers, trees, and the abiotic parts of the environment have been the subject of a fight for rights because humans have decided to become their champions (Stone, 1972).

The latest installment in the saga of rights–the fight for other's rights–are robots. While robots are somewhat new, the debates they give rise to are arguably not, as they draw upon and continue debates from environmental ethics. While not new, the question of how artificial entities fit into these old debates is attracting increased attention (Harris and Anthis, 2021). Old arguments, on old battlegrounds, are rehashed, as robot champions (champions for the rights of robots) clash with those who call the fight for robots right ludicrous, unthinkable, or just outright harmful and disruptive of the fight for equal get for all humans (Birhane and Van Dijk, 2020a; 2020b). The latter group is by some branded human chauvinists since their arguments are considered to be anthropocentric, and they are consequently criticized and labeled as both backward, unjust, and ignorant of history.

One particular form of argument for imagining robot rights is relationalism, with Coeckelbergh. (2010), Coeckelbergh. (2011), Jones. (2013), Gunkel. (2018b), and Gellers. (2020) as some of its champions in arms. Thinking otherwise, Gunkel calls it, when he argues that relational ethics opens the door for seriously considering robot rights and taking a step or two past anthropocentrism. In this article, I challenge the implicit and at times explicit claim that relationalism allows us to move past anthropocentrism, as I argue that the approach is in fact a form of neo-anthropocentrism that recenters human beings and their unique ontological properties, perceptions, and values and that this is quite the opposite of the stated purpose of this purported thinking outside the box (Gunkel, 2018b). I do so by raising three objections: 1) relationalism centers human values and perspectives, 2) it is indirectly a type of properties-based approach, and 3) edge cases reveal potentially absurd implications in practice.

My goal is thus to challenge the proponents of this approach to clarify and further develop their theories, and others have similarly claimed that relationism "leaves us with many unresolved questions" (Tavani, 2018). I will, however, not pursue the question of whether or not the relational approach is more useful or better than the alternatives, as the purpose is to highlight issues related to the anthropocentric nature of the approach. This also means that I will not be evaluating the different varieties and the nuances of the various philosophical foundations used by the different researchers in this tradition, beyond what is required for establishing whether or not the emerging tradition–as a tradition–is anthropocentric.

In order to evaluate the nature of relationalism, in *Anthropocentrism and the Others* I examine the nature of anthropocentrism and non-anthropocentrism. More importantly, I highlight the importance of examining the different types of each, because the umbrella terms "anthropocentrism" and "non-anthropocentrism" themselves contain too much variation to be philosophically meaningful. In *The relational Turn as Neo-Anthropocentrism* I move on to relationalism, to briefly present how its proponents present the approach before I proceed to examine it in light of the types discussed in *Anthropocentrism and the Others*. I end this section by presenting the three objections which together constitute my challenges to relationalism.

ANTHROPOCENTRISM AND THE OTHERS

As the starting point for the examination of the nature of the relational turn in the robot rights discourse, the field of environmental ethics provides a range of applicable tools and concepts. It could even be argued that the question of the moral standing of robots is a part of environmental ethics, and does not necessitate new forms of ethics such as robot ethics. Environmental ethics is, after all, at times understood as the examination of how moral thinking and action can be expanded both beyond humans and beyond the present (Nolt, 2014). Just as the robot rights movement is often perceived as a form of unwarranted and misdirected activism (Birhane and Van Dijk, 2020a; Birhane and Van Dijk, 2020b), the same often goes for environmental ethicists, at times labeled "treehuggers," anti-humanists or misanthropes who fight for the rights of animals and the natural world at the expense of human beings (Drengson, 1995; Kopnina et al., 2018; Rottman et al., 2021). Such a denouncement is, however, based on the erroneous notion that there is a "hierarchy of ethics" and that all research should be directed to whichever problems the critics consider to be more important than considering robot–or environmental–rights (Sætra and Fosch-Villaronga, 2021).

Key concepts related to the moral standing of robots are moral community, moral agency, and moral patiency (Nolt, 2014). All entities that are deemed worthy of moral consideration belong to the moral community, and anyone who has a claim on moral consideration is a moral patient. Some entities will have such a claim and an associated moral duty, and these are considered moral agents. The moral community is here considered a purely hypothetical construct, and any type of moral community is theoretically possible. One theory might argue that only humans have a claim for moral consideration, while another might argue that humans are not even parts of the moral community.

Robots can, in theory, certainly be considered parts of our moral community, but few as of yet have argued that they are full-fledged moral agents. Most existing theories will consider humans moral agents while including some other entities as moral patients with various claims to moral consideration. It is important to stress that no universally accepted definitions of which traits warrant moral patiency exist (Gellers, 2020; Gunkel, 2018b), and this is one of the very reasons for the emergence of the relational approach, as we will later see. Neither are the criteria for moral agency sufficiently clear to serve as the basis of agreement between moral philosophers of various stripes. However, I'll argue that it is reasonable to posit that if we are to provide an entity with a duty to consider the moral claims of others, they must have a least the semblance of the sensory and cognitive capacities to do so–they must have moral competence (Nolt, 2014). As argued by Næss. (1989) humans are the first species with the capacity to understand how their behavior affects other beings and consequently change this behavior to achieve some form of equilibrium. Humans can, he argues, "perceive and care for the diversity of their surroundings" (Næss, 1989, p. 23), even if they arguably do not always do so.

This framework allows us to examine how different people ascribe moral standing to different forms of entities, potentially including robots. Rather than focusing on the resulting ascription of moral standing to various entities, I'm concerned with how the theorists most clearly associated with the "relational turn" in robot ethics arrive at the possibility of rights for robots, and in particular whether or their approach is less anthropocentric than other approaches. In order to achieve this, a somewhat roundabout trip into the murky definitional waters surrounding anthropocentrism and non-anthropocentrism is required. That is because, as we shall see, these terms are often used in a confusing and non-specific manner.

Anthropocentrism

Ethical theories that assume the centrality of humans in any consideration of moral standing or value are often referred to as anthropocentric (Nolt, 2014). Posthumanism–focusing on the "decentering of humans"–and the biospherical egalitarianism of Næss, for example, might both be argued to entail clear rejections of anthropocentrism (Næss, 1973; Meyer, 2001; Braidotti, 2013). Robot rights is another phenomenon that might at first seem to be–and is often argued to be–non-anthropocentric. Concluding that they are non-anthropocentric is, however, premature, as I will argue that even theories that do not consider humans to be most–or the only things–valuable can clearly be anthropocentric, depending on their method of ascribing moral standing to others. To make this point, the most important types of anthropocentrism must be examined.[1] I rely on Nolt (2014) terminology for distinguishing between axiological and ethical anthropocentrism in what follows.

Axiological Anthropocentrism

While there may be many reasons for humans putting themselves at the center of any moral examination, Nolt. (2014) points to the emergence of monotheism as the key factor which led to a focus on humanity as a meaningful group, rather than mere humans as contained in distinct cultures and smaller groups. He then connects anthropocentrism and monotheism to the notion of human rights, which is arguably one of the most important manifestations of anthropocentric ethics prevalent in modern societies (Zimmerman, 1985). Much like Aristotle had done before them, late antiquity Western philosophers, working in the age of emerging monotheism, saw the world as "an exquisitely designed hierarchical structure in which all things had God-given values and purposes" (Nolt, 2014, p. 64). All things, according to this line of thinking, exist to serve the needs of "higher" entities. Water serves the needs of plants, which serve the need of animals, which serve the need of humans, for example. A central example of this line of thinking is the idea of a great chain of being (Lovejoy, 1936). In the original chain angels and God superseded humans, while the modern secular and anthropocentric version could be argued to place humans at the pinnacle.

This kind of anthropocentrism, with humans at the center and everything else assigned value by how it serves human needs, is referred to as axiological anthropocentrism (Nolt, 2014). If humans, for example, tend to see something of themselves–something that they value–in robots, this could become the basis of offering such entities some form of moral consideration, or protection (Darling, 2016a). This also relates to the notion that how we treat other entities impacts us. If, for example, humans somehow hurt themselves by mistreating animals, or robots (Darling, 2016a), this could give rise to both moral and legal protection of such entities. Not for the entitie's sake, but for ours. The moral standing and value of entities are, according to this theory, assigned by and for human purposes. For us, by us. While somewhat similar to the religious view that things have value according to God's desires and purposes, it is clearly distinct, as no God or gods serve any necessary purpose in axiological anthropocentrism. If humans find a purpose for God, however, God will be valued accordingly, but not the other way around.

Ethical Anthropocentrism

A different form of anthropocentrism, one that is often conflated with the axiological variety, is ethical anthropocentrism. This is a theory that encompasses the view that humans are morally considerable, and most other things are not (Nolt, 2014). The strict variety states that only humans are morally considerable, while other varieties assign some–but not much–value to some non-human entities. It is easy to see why axiological and ethical anthropocentrism are often conflated, but it is still important to distinguish between them. Axiological anthropocentrism allows for a far more inclusive moral community than ethical anthropocentrism, provided that humans find value in other beings, or that humans simply find value in providing other beings with moral standing. The common denominator is that humans are centered, as they are either of superior moral standing or the only cause of moral standing provided to others. The instrumental approach to robot rights is often based on ethical anthropocentrism, and, for example, the notion that robots should be slaves (Bryson, 2010), is based on the idea that they do not have moral worth, even if they can indeed be useful. According to such positions, they could be said to have instrumental value to humans, but no intrinsic value. Ethical anthropocentrism is also clearly linked to the idea of human chauvinism, which is, according to some, deeply embedded in western culture and consciousness (Seed, 1988).

Non-Anthropocentrism

While we are currently seeing a growing concern for the environment–in the shape of animals, the climate, or various ecosystems–it is still considered relatively radical to argue in favor of non-anthropocentric ethical theories. After all, a whole lot of those concerned with the climate and biological diversity, for example, make few efforts to hide the fact that their concerns stem from the negative effects for human beings if the environment is impoverished or changed in ways unfavorable to human flourishing. Truly non-anthropocentric theories must argue in favor of the worth of nature regardless of how nature impacts humans, and as with anthropocentrism, there are several types of non-anthropocentrism.

Rights and Ontology-Based Ethical Non-Anthropocentrism

One apparent way to move past anthropocentrism is to provide other entities with rights. I will only consider the approach that assumes that the entities that receive rights are capable of being bearers of rights, and not an entirely legalistic approach that ascribes "rights" to nature, corporations, etc., for merely instrumental purposes. Robots can, Sætra (2021a) argues, certainly be considered as some sort of limited liability

[1] One distinction I will not pursue in this article is that between short- and long-term anthropocentrism, which distinguishes between those that believe only humans that live right now have moral standing and those that consider potential future humans as well

corporations if this serves socio-political needs, but this must not be conflated with the notion that legal status also provides moral status.

Peter Singer and Tom Reagan and Singer. (1976) are two well-known proponents of animal rights, but others have also asked whether, for example, trees, rivers, or entire ecosystems have rights (Stone, 1972). The question given rise to by all these approaches is, however: what are rights derived from? Answers range from philosophy, deities, natural law, the use of (human) reason, etc. One particular approach, which is the loci of this article, is one where rights are ascribed on the basis of the nature of relationships, and the examination of this approach will be saved for the next chapter.

One kind of non-anthropocentrism uses traits, or ontological features of entities, to argue that humans are not really special, and thus do not deserve a special moral status. These theories are often related to the problems of demarcation that arise as soon as someone attempts to defend ethical anthropocentrism by describing why humans are different from animals (Nolt, 2014). In a discussion of what distinguishes humans from machines, for example, Sætra (2019) examines a range of different properties, or traits, that have historically been used to distinguish humans from other entities. Reason, a soul, life, etc.–all these concepts fail, he argues, as the basis for clearly demarcating humans from others. When traits are tested as criteria for human value, marginal cases are often used to demonstrate the problems associated with the various criteria (Dombrowski, 1997; Nolt, 2014). For example, if reason is our criteria, how do we deal with the fact that some animals have more of it than some humans (Sætra, 2019)? Such a traits-based approach could arguably be both anthropocentric and non-anthropocentric, and also anthropocentric in two different ways. If traits are chosen in order to limit moral consideration to *Homo sapiens,* we get ethical anthropocentrism, while other approaches focus on traits shared by other entities as well–such as sentience–in which we might have axiological anthropocentrism.

Another way to potentially derive rights, obligations, and moral standing from a situation in which none exist is the contractualist approach (Hobbes, 1946). With this approach, rights are derived from consent based on contract, but neither the contract nor the consent need be explicit. In Hobbe's social contract theory, for example, the contract can be considered a hypothetical thought experiment aimed at generating agreement of what reasonable people would agree to, and thus the contract is not taken to be an actual contract that each and every individual has actually agreed to (Sætra, 2014). While Sætra. (2014) argues that contractualism might lead to a type of environmentalism based on human self-interest, the social contract theorist Carruthers. (1992) has warned that if we extend our moral communities to encompass other entities, morality might be diluted. Despite such warnings, contractualism can potentially lead to rights for others, just as we have extended rights to animals and a range of other beings that cannot themselves be an active contracting party. One approach to a contractual approach to non-human rights is to have humans serve as curators or guardians (Sætra, 2014).

The notion of rights is a topic worthy of its own article, and as I am mainly concerned with understanding how the relational approach to moral standing relates to anthropomorphism, I save the rest of this discussion for *The relational Turn as Neo-Anthropocentrism*, in which the relational turn is examined in more detail. Before that, two non-anthropocentric varieties will provide more insight into just how we might justify the decentering of humans.

Axiological Non-Anthropocentrism

Axiological anthropocentrism ascribes value to entities according to how valuable they are perceived to be by humans. Humans are a relatively diverse bunch, however, and it is here important to be wary of the danger of conflating western values with human values (Gellers, 2020). Moving beyond the discussion of whose human values matter, the axiologically non-anthropocentric approach starts with the assumption that human values are not the only values (Nolt, 2014). Other entities might indeed have values as well, or things might at the very least potentially be good or bad for them. The problem with this theory, as compared to the anthropocentric variety, is that it is difficult to determine what the values of non-human entities really are. Three main approaches to discovering these are the hedonistic, preference-satisfaction, and objective welfare, all with their distinct strengths and obvious weaknesses (Nolt, 2014).

The hedonistic approach entails an emphasis on aggregate pleasure and pain, and it is often associated with the consequentialist variety of ethics referred to as utilitarianism. However, concerning moral standing, the ability to experience pleasure and pain is what matters, and sentientism is perhaps the most prominent variety in this category. The key objection to sentientism in the context of robot rights is that it once again entails examining the ontological status of subjects–here the capability of sentience. Furthermore, since it is most often biocentric (Nolt, 2014), it tends to exclude machines. However, critics argue that it is difficult to distinguish human pain and pleasure from what is "experienced" by a sophisticated machine, just as reason and other objective qualifiers also bring us into murky waters. This objection to biocentrism becomes increasingly relevant with modern advances in biomimetic robots (Winfield, 2012), and various robots built to model human emotions, homeostasis etc. (Cominelli et al., 2018; Man and Damasio, 2019).

Preference-satisfaction is a broader form of consequentialism in which entities may be thought to have interests beyond pleasure and pain, and what subjects themselves consider good and desire is what matters. But how do we uncover the preferences of entities that cannot speak or express themselves? Those who are capable of acting are helpful in that we might propose using the theory of revealed preferences from economic theory (Samuelson, 1948), but what about abiotic nature? And what about robots, who can both speak and act? This is where the question of agency comes up, and in this article, I adhere to the position that robots cannot as of now be said to be capable of owning and being responsible for their own actions, and consequently, I assume that their words and actions do not represent the robot's own preferences in a meaningful way (Sætra, 2021a).

The final approach consists of basing one's evaluation on some notion of objective welfare–an approach that might result from wanting to deduce what is good for, and thus assumed preferred, by entities. Næss (1989), for example, uses the notion of flourishing as a fundamental good for all entities. Highly useful for dealing with entities that neither speak nor act, but also for humans who might not realize their own best interests. Or at least so the paternalists might say.

Ethical Biocentrism

A different approach is one in which entities are quite simply regarded as origins of value. Inherent, or intrinsic, value is the term often used for this approach (Næss, 1989; Nolt, 2014), as their value is assumed to be entirely disconnected from their instrumental value to humans or to how humans imagine value. One example of such a theory is Arne Næss's deep ecology, based on the notion of biocentric egalitarianism and an outright rejection of anthropocentrism and the superior moral value of humans (Næss, 1989).

As compared to the previous type of non-anthropocentrism, ethical biocentrism does not require us to uncover, or conjure up, the interests, preferences, etc., of other entities. Instead, they are considered valuable just because of being what they are, which is why the terms intrinsic or inherent value are often used. Midges and ticks, for example, have very little going for them in terms of obvious instrumental value for humans–particularly if people's opinions about them, rather than the ecosystem services they provide–are the basis of ascribing value. However, in theories such as Næss's deep ecology, even such beings are considered valuable by virtue of simply being what they are, and they are provided with the same rights to flourish like the rest of us. The details of what constitutes inherent or intrinsic value, and the differences used in describing who and what has such value, is beyond the scope of this article, and it suffices for now state that such approaches are effective in ascribing rights to animals, and at other parts of nature, while it has not been particularly useful for imagining robot rights.

THE RELATIONAL TURN AS NEO-ANTHROPOCENTRISM

The time has come to consider how to categorize theories belonging to what is often referred to as the "relational turn"[2] in robot ethics (Gerdes, 2016). The goal of what follows is not to examine whether or not "relationalism" (Coeckelbergh, 2010), "social-relational ethics" (Coeckelbergh, 2010; Gunkel, 2018b; Harris and Anthis, 2021), or "ecological relationalism" (Jones, 2013) is right, wrong, beneficial, etc. Neither is it a deep philosophical examination of the nuances and differences between the various manifestations of relationalism beyond what is required to establish the fundamental approach shared by these theorists.

Rather, the goal is to examine whether relationalism is anthropocentric or non-anthropocentric, and using the terms established also determines more specifically which type most accurately describes it. The reason why uncovering this is of interest is that it is relevant to the discussions that emerge as soon as differences of opinion with regard to the possibility or desirability of robot rights surface. In such discussions, opponents of robot rights might argue that pursuing such questions is pointless, or even outright harmful, as more important questions related to human flourishing should be prioritized (Birhane and van Dijk, 2020b). Some arguing in favor of robot rights might then accuse the opponent of being human chauvinists, and either explicitly or implicitly indicate that the opponents are anthropocentric human chauvinists, while those open to robot rights are not. Whether or not these proponents are right is what I address in the following.

The Relational Turn

What is here described as the relational turn refers to the idea that moral consideration should be premised on social relations rather than ontological or socio-political frameworks (Coeckelbergh, 2010). What I refer to as relationism is not the particular philosophy of one person, however, and I will, in general, refer to relationism as a tradition manifested through the work of Mark Coeckelbergh. (2010) Raya Jones. (2013), David Gunkel. (2018b), and Josh Gellers. (2020).

Relationalism is, however, not a new phenomenon, and it is often traced back to the relational approach of Arne Næss (Brennan and Lo, 2021; Næss, 1989). It is also closely related to care ethics (Donovan and Adams, 1996, Donovan and Adams, 2007), which emphasizes relationships, and both anthropocentric and non-anthropocentric varieties of care ethics have been proposed. Common for the traditional care ethics is that they are routinely criticized for their inability to extend rights to strangers–both humans and other types of others (Nolt, 2014).

One response to this is the relational ethic of Palmer. (2010), which is based on ecofeminism and its emphasis on relationships rather than individuals in isolation (Palmer, 2003). She, like the relationists in the robot ethics camp, suggests that responsibilities and moral standing are not just matters of capabilities, but also of our interactions with others. However, relationships are used differently by Palmer than by the robot ethicists, as she argues that actual interactions create responsibilities, more so than relatively abstract notions about how humans, in general, might be capable of forming relations with other entities. Consequently, I focus on relationalism as it is detailed in the robot ethics discourse, as robot relationalism and the traditional varieties just discussed are somewhat different.

The problem with traditional theories, Coeckelbergh. (2010) argues, is that they all–deontological, utilitarian, and partly virtue ethics–rely on what he calls "ontological features" of the entities in question. These are, for example, requirements related to

[2] In the following, I will mainly refer to these theories as relationalism, unless specifically pointing to the various names the different authors themselves use. It is also worth noting that it seems unfortunate to speak of these theories as "relational ethic," as that is an already-existing field of study focusing on ethical conduct in various relationships, such as nurse-patient relationships (Ellis, 2007).

biological life, rationality, sentience, etc., as we have already seen. He proceeds to argue that we need a "social ecology," which is–much like Arne Næss deep ecology–based on the science of ecology combined with "Eastern worldviews." This is also discussed at some length by Næss himself, and Jones. (2013), Jones. (2015), Gunkel. (2018b), Gellers. (2020), all emphasize the need for and potential utility of moving beyond traditional Western world-views in order to arrive at an improved understanding of how to understand moral standing and the nature of various others.

What becomes important, then, is the network of interactions and relations between entities, and not the entities in isolation and their properties.

> The alternative approach I propose attempts to avoid the skepticism by replacing the requirement that we have certain knowledge about real ontological features of the entity by the requirement that we experience the features of the entity as they appear to us in the context of the concrete human-robot relations and the wider social structures in which that relation is embedded (Coeckelbergh, 2010, p. 14).

Rather than ascribing moral standing on the basis of characteristics of the entity–the properties-based approach–the very fact that we relate with other entities becomes the basis for obligations and claim for moral consideration (Gellers, 2020). As with the traditional types of relational theories, what matters is not necessarily whether or not the others are like us (Darling, 2016a), or if we see ourselves reflected in them (Sætra, 2021b), but rather how these others become actors in our social structures with which we interact. What I refer to as relationalism is, as mentioned, often referred to as social-relational ethics for this reason. This is also related to Arne Næss's notion that self-realization entails coming to see ourselves as nodes in a relational total-field image (Næss, 1989). This leads to identifying with the other nodes in the field, and this is in theory accompanied by an acknowledgment of how all is, in reality, one, and all of value. While deep ecology opens for including both biotic and abiotic nature in this field, artificial life has no obvious function in this network. In robot relationalism, the fact that we relate to robots is taken as an indication that these entities are, in fact, nodes of value due to these relations.

The key argument, as presented by Gunkel. (2018b), is that our evaluation of what another entity is, in a moral context, depends on how it is treated and not some isolated consideration of what the thing in isolation is. In particular, it is important that the other is not simply reduced to a reflection of ourselves and some sort of alter-ego which is perceived as valuable because of its likeness to us (Gunkel, 2018a). Gunkel draws heavily on Levinasian philosophy in his explorations of the potential for robot rights, and it is interesting to note how he argues that it is important to "break free from the gravitational pull of Levinas's own anthropocentric interpretations" (Gunkel, 2018a, p. 97). In his reinterpretation of Levina's philosophy in a way that opens for considering robots to be meaningful others–as Levinas himself does not do–the question is whether Gunkel simply develops a new kind of anthropocentric theory or if he arrives at a non-anthropocentric theory. The objections which are shortly presented suggest that what has occurred is a move from ethical to axiological anthropocentrism and not a move to non-anthropocentrism.

Gellers. (2020) explores both the legal and moral status of robots, and while the former is outside the scope of this article, his considerations regarding the latter are largely in line with the preceding authors. He explicitly argues that the relational ethic proposed by Levinasian scholars (and he here includes Coeckelbergh and Gunkel) is both promising and that it has moved old debates, but also that it might go too far in abandoning the role of properties. This–the role of properties in our encounters with others–is also the very basis of one of my challenges.

What is of most interest here is not the nuances of the different varieties of relationalism, but whether or not relationalism really succeeds in moving us past anthropocentrism, or if it is instead a new type of anthropocentrism.

Neo-Anthropocentric Relationalism

On the basis of the preceding considerations, I now turn to an explanation of why I argue that relationalism is a neo-anthropocentric ethical theory. Rather than providing the means to move beyond anthropocentrism and the traits-based approach, I argue, based on the objections presented below, that the theory is both anthropocentric and dependent on traits-based considerations. In addition, the theory is faced with a practical challenge related to its potential use in practice. In the following I outline three main challenges for relationalism: 1) relationalism centers human values and perspectives, 2) it is indirectly a type of properties-based approach, and 3) edge cases reveal potentially absurd implications in practice.

Human-Centered Relations

My first objection is that relationalism is arguably deeply anthropocentric because moral standing is derived exclusively from how human beings perceive and form relations with other entities. As we have seen, moral standing is here derived from how something is treated, and not what it is. This means that humans are key to determining value, as it is how entities are treated and perceived by humans that determine their moral standing (Gunkel, 2018b). While this surely opens the door for moral standing for robots that are able to mobilize human social instincts and trigger social responses (Sætra, 2020), it is hard to see how this constitutes a form of non-anthropocentrism. On the contrary, it seems like a clear representation of a system based on the axiological anthropocentrism defined in *Axiological Anthropocentrism*. It is interesting to note that I here levy the same kind of criticism against relationalism that Gunkel (2018a, p. 95) uses as an objection against Darling. (2016a): "because what ultimately matters is how "we" see things, this proposal remains thoroughly anthropocentric and instrumentalizes others." While Gunkel imagines the other as something more than a mirror-image, using Levina's theory to modify our understanding of the other arguably does not introduce reciprocity or a true recognition of the other for their own sake, since it is human perceptions and experience of the other that is used as the basis for determining value. Thus, relationalism is subject to the very same critique aimed by its proponent on another theory.

Two different anthropocentric doctrines could be developed from relationalism. One is similar to the relationism of Palmer. (2010), in which relations with actual entities are constitutive of moral responsibilities. In the case of robots, I might interact with a Paro robotic seal (Paro Robots, 2021), for example, and feel that I have developed a certain rapport with this entity. This would in turn create responsibilities towards that particular robot, but not toward other Paros. The other approach would be to argue on a more detached level that actual relations are not relevant, while the potential to form relations is Such a doctrine would entail that if humans are capable of forming relationships with Paro robots, then all Paro robots must be awarded moral consideration. This, however, would take us right back to the properties-based approach that the relationists purportedly want to move past.

Of central importance, however, is the fact that the relations being described by the robot ethicists are arguably not really based on true relations at all, as the emphasis is not on mutuality but on how humans perceive and treat other entities. What occurs in the other entity is seemingly irrelevant, and further highlights the relatively extreme anthropocentric nature of the theory. It must be noted that this stands in contrast to care ethics and relational ethics as established in the domain of animal ethics, in which mutuality is a fundamental part of any relationship worth considering. In robot relationism, mutuality and considerations regarding the capabilities, intentions, experiences, etc., of the other is excluded from the analysis, and this leaves us with a peculiar one-sided approach to relations that gives rise to my challenges.

Also of importance is the fact that anthropocentrism is not necessarily a bad thing, once properly understood. Non-anthropocentrism is a term mired in difficulty, as some argue that it is impossible for humans to avoid being anthropocentric, as the very notion of value–either instrumental or intrinsic–is necessarily based on a human perspective (Hayward, 1997; Hargrove, 2003). If robot rights theorists accept this view, however, they might be better off by using Hargrove's term weak anthropocentrism (2003) to describe their own theory, and argue why this is preferable to strong anthropocentrism. This would dramatically clarify these debates and would be an improvement over a situation in which anthropocentrism alone is assumed to provide sufficient clarity. It does not, and consequently needs to be further elaborated.

Properties Strike Back

My second objection is that relationalism is in reality a camouflaged variety of the properties-based approach. This is so because how we relate to other entities is determined by the properties of these others. At the very least, as Gellers. (2020) acknowledges, it significantly influences relations. However, we cannot arguably perceive someone's true nature, intentions, feelings, etc., so how are the perceived relations with others arrived at?

As discussed in the previous challenge, proponents of relationism generally tend to argue that we need not consider the "internal" properties of robots. If a robot acts in ways that allow it to engage in the kinds of social interactions with humans that the relationists deem important, this is sufficient (Tavani, 2018). This, again, relates to what Danaher (2020) calls ethical behaviorism, which entails that moral duties and responsibilities are grounded in external and observable action, and not entitie's internal processes and mechanisms.

However, how we relate to someone, and how an entity acts, is dependent on their properties. I might, for example, say that I do not care what species something is, but will evaluate moral standing merely by how I relate to it. The problem, then, is that this will often entail providing moral standing to exactly the same entities as before because those with the properties of humans are the ones I relate to in the manner I consider to be constitutive of moral standing.

It is easy to see why relationalism has emerged so clearly in the discourse on robot rights, as robots are now designed with a range of exactly those properties that are conducive to social relations (Sætra, 2020). It is also an approach that takes us past what might be labeled biological chauvinism, as traditional theories, both anthropocentric and non-anthropocentric, have often focused on the biological foundation of life and moral standing (Gellers, 2020; Manzotti and Jeschke, 2016). Emphasizing biology is indeed problematic, as it excludes mechanical robots from consideration, while it also introduces problems related to the status of humans who integrate with non-biological technology (Sætra, 2019).

Once again, this creates the foundation for two different strands of relationalism. One in which actual properties and capacities required for mutual and reciprocated relationships are used as the basis of determining the potential for relationships, and one in which perceived properties are taken into consideration. The latter strand gives rise to the third objection described below, while the former arguably excludes even the most sophisticated biomimetic robots as parties to relationships.

Relationalism is at times also argued to be able to account for changing social relations and social constructs (Gunkel, 2018b), and this is perceived as an advantage of the approach. In response, proponents of properties-based theories could point out that this is also the case for traditional properties-based approaches, as phenomena like rationality, sentience, consciousness, etc., are also social constructs that change over time, with clear consequences for who and what are accorded moral worth.

Edges Cases and the Problem of Anthropomorphism

My third challenge is that any approach to moral standing based on one-sided "relations" based exclusively on perceived properties, coupled with the human tendency to anthropomorphize other entities, leads to potentially absurd implications when the theory is applied in practice. Anthropomorphism describes the process of attributing human properties to other things, i.e. robots, and this can occur intentionally or unintentionally (Coeckelbergh, 2021). The situation that ensues is one in which humans might anthropomorphize other entities, and consequently feel that they relate to these things, which in turn triggers the relationist inclination to use this to accord the thing moral standing.

People anthropomorphize social robots and tend to attribute various traits such as purpose, intentions, etc. to them, despite these robots not actually having such traits of capabilities (Sætra, 2021a). But people also anthropomorphize a wide range of less sophisticated things. A volleyball, for example, in the movie Cast Away, but also a wide range of other things, such as computers, dolls, etc. (Reeves and Nass, 1996; Levy, 2009; Darling, 2016b). In such situations, who is to decide whether these relations, which people according to their own subjective experience are forming with these things, are constitutive of moral standing for the non-human part of the relationship or not?

Are we to rely on objective evaluations of which relationships should matter? If so, the process and criteria by which to perform such evaluations–and who will perform them–become important questions that provide the ground for much potential conflict. On the other hand, if subjective evaluations are to be given consideration, we might in fact end up with a form of subjective relationism where robots are provided moral standing. But so is potentially a volleyball, and a child's security blanket. As a basis for arriving at a universal theory for determining moral standing, an approach with arguably absurd implications seem to require some more development before it is workable. At the very least, it highlights how an ethical theory of moral standing such as relationalism does not give rise to objective and universal outcomes, but rather show the political nature of deciding who in the end decides.

CONCLUSION

The question I have sought to answer is whether relationalism can help move us past chauvinistic and anthropocentric moral theories. I have accepted that relationalism can indeed be effective in allowing non-humans to be awarded moral standing, but I have also argued that the method by which it does so is beleaguered by certain fundamental problems. Firstly, that it is anthropocentric, much like the theories it seeks to replace. Secondly, that it is based on traits, either objectively or subjectively assessed. Thirdly, that anthropomorphism potentially leads to absurdities whenever relationism is used as the basis of determining moral standing without combining relationism with a properties-based approach.

Relationalism is, I argue, based on *Axiological Anthropocentrism*. However, as the theory is not premised on the explicit centering of human perception of value as the basis of moral standing, and since it is also proclaimed to be a solution to the problems of both traditional anthropocentric and non-anthropocentric theories, the theory will be labeled neo-anthropocentric–a type of new anthropocentrism. The novelty of the theory is that it dismisses all explicit references to the superior moral status of humans or human instrumental value as the basis of moral valuation, which means that it is not based on ethical anthropocentrism and human chauvinism. However, it also rejects non-anthropocentric theorie's adherence to concepts such as inherent value or biocentric egalitarianism. Relationism, then, might not take us past anthropocentrism, but it does take us past human chauvinism. I have also suggested that its proponents might highlight the superiority of weak over strong anthropocentrism (Hargrove, 2003), and Hayward. (1997) has similarly argued that it is not anthropocentrism itself which is the problem, but the various forms of speciesism and human chauvinism.

Relationist neo-anthropocentrism allows us to explore the potential of social ecology as the basis for determining moral standing, and this is indeed valuable, as also shown through traditional approaches to relational ethics and care ethics. However, in contrast to relational non-anthropocentric approaches, such as deep ecology, this form of social ecology explicitly centers humans as their treatment of and relations with others give rise to the other's moral standing. Relationalism as it is used in the robot ethics discourse provides an interesting theoretical path towards providing robots with moral standing. It is, however, beleaguered by a number of challenges, and this article is intended as a challenge and a request for further elaboration of the approach and theories based on this approach in order to clarify and more clearly position the theory in relation to other related theories and concepts from environmental ethics, such as anthropocentrism. This is not to say that proponents of a relational ethics have not acknowledged the challenges and complexities associated with relationalism–they have–but merely to state that there are still questions that need answering and clarifications that must be made.

On a closing note, some proponents of relationalism might come to accept the label of neo-anthropocentrism and their reliance on a traits-based approach. However, a consequence of this would be that some of the purported advantages of relationism–such as removing us from human chauvinism and the problematic focus on traits–would have to be abandoned. Gellers (2020, p. 153) has to some extent done just this, as he argues that his "explicitly relational" approach must to some extent be combined with the properties-based approach, even if this reintroduces some of the problems associated with properties as a basis of moral standing. And, while it may not be chauvinistic, it is anthropocentric.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Birhane, A., and Van Dijk, J. (2020a). A Misdirected Application of AI Ethics. Noema. Available at: https://www.noemamag.com/a-misdirected-application-of-ai-ethics/ (Accessed July 15, 2021).

Birhane, A., and van Dijk, J. (2020b). "Robot Rights? Let's Talk about Human Welfare Instead," in Paper presented at the Proceedings of the AAAI/ ACM Conference on AI, Virtual: February 2–9, 2021, (Ethics, and Society).

Braidotti, R. (2013). Posthuman Humanities. *Eur. Educ. Res. J.* 12 (1), 1–19. doi:10.2304/eerj.2013.12.1.1

Brennan, A., and Lo, Y.-S. (2021). "Environmental Ethics," in *The Stanford Encyclopedia of Philosophy*. Editor E. N. Zalta Summer 2021 Edition ed. plato.stanford.edu. Available at https://plato.stanford.edu/archives/sum2021/entries/ethics-environmental/.

Bryson, J. J. (2010). "Robots Should Be Slaves," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. Editor Y. Wilks (Amsterdam: John Benjamin), 63–74. doi:10.1075/nlp.8.11bry

Carruthers, P. (1992). *The Animals Issue: Moral Theory in Practice*. Cambridge: Cambridge University Press.

Coeckelbergh, M. (2011). Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. *Int. J. Soc. Robotics* 3 (2), 197–204. doi:10.1007/s12369-010-0075-6

Coeckelbergh, M. (2010). Robot Rights? towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12 (3), 209–221. doi:10.1007/s10676-010-9235-5

Coeckelbergh, M. (2021). Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach. *Int. J. Soc. Robotics*, 1–13. doi:10.1007/s12369-021-00770-0

Cominelli, L., Mazzei, D., and De Rossi, D. E. (2018). SEAI: Social Emotional Artificial Intelligence Based on Damasio's Theory of Mind. *Front. Robot. AI* 5, 6. doi:10.3389/frobt.2018.00006

Danaher, J. (2020). Welcoming Robots into the Moral circle: A Defence of Ethical Behaviourism. *Sci. Eng. Ethics* 26 (4), 2023–2049. doi:10.1007/s11948-019-00119-x

Darling, K. (2016b). "'Who's Johnny?'Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy," in *ROBOT ETHICS 2.0*. Editors G. B. P. Lin, K. Abney, and R. Jenkins (Oxford: Oxford University Press).

Darling, K. (2016a). "Extending Legal protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects," in *Robot Law*. Editors R. Calo, A. M. Froomkin, and I. Kerr (MA: Edward Elgar Publishing), 213–231.

Dombrowski, D. A. (1997). *Babies and Beasts: The Argument from Marginal Cases*. Champaign: University of Illinois Press.

Donovan, J., and Adams, C. J. (1996). *Beyond Animal Rights: A Feminist Caring Ethic for the Treatment of Animals*. New York: Continuum Intl Pub Group.

Donovan, J., and Adams, C. J. (2007). *The Feminist Care Tradition in Animal Ethics: A Reader*. Columbia University Press.

Drengson, A. (1995). The Deep Ecology Movement. *The Trumpeter* 12 (3).

Ellis, C. (2007). Telling Secrets, Revealing Lives. *Qual. Inq.* 13 (1), 3–29. doi:10.1177/1077800406294947

Gellers, J. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. Abingdon: Routledge.

Gerdes, A. (2016). The Issue of Moral Consideration in Robot Ethics. *SIGCAS Comput. Soc.* 45 (3), 274–279. doi:10.1145/2874239.2874278

Gunkel, D. J. (2018b). *Robot Rights*. London: MIT Press.

Gunkel, D. J. (2018a). The Other Question: Can and Should Robots Have Rights?. *Ethics Inf. Technol.* 20 (2), 87–99. doi:10.1007/s10676-017-9442-4

Hargrove, E. (2003). "Weak Anthropocentric Intrinsic Value," in *Environmental Ethics: An Nthrology*. Editors A. Light and H. RolstonIII (Malden: Blackwell).

Harris, J., and Anthis, J. R. (2021). The Moral Consideration of Artificial Entities: A Literature Review. *Sci. Eng. Ethics* 27, 53. doi:10.1007/s11948-021-00331-8

Hayward, T. (1997). Anthropocentrism: A Misunderstood Problem. *Environ. Values* 6 (1), 49–63. doi:10.3197/096327197776679185

Hobbes, T. (1946). *Leviathan*. London: Basil Blackwell.

Jones, R. A. (2015). *Personhood and Social Robotics: A Psychological Consideration*. London: Routledge.

Jones, R. A. (2013). Relationalism through Social Robotics. *J. Theor. Soc Behav* 43 (4), 405–424. doi:10.1111/jtsb.12016

Kopnina, H., Washington, H., Taylor, B., and J Piccolo, J. (2018). Anthropocentrism: More Than Just a Misunderstood Problem. *J. Agric. Environ. Ethics* 31 (1), 109–127. doi:10.1007/s10806-018-9711-1

Levy, D. (2009). *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper Collins e-books.

Lovejoy, A. O. (1936). *The Great Chain of Being*. Cambridge: Harvard University Press.

Man, K., and Damasio, A. (2019). Homeostasis and Soft Robotics in the Design of Feeling Machines. *Nat. Mach Intell.* 1 (10), 446–452. doi:10.1038/s42256-019-0103-7

Manzotti, R., and Jeschke, S. (2016). A Causal Foundation for Consciousness in Biological and Artificial Agents. *Cogn. Syst. Res.* 40, 172–185. doi:10.1016/j.cogsys.2015.11.001

Meyer, J. M. (2001). *Political Nature: Environmentalism and the Interpretation of Western Thought*. Cambridge: Mit Press.

Næss, A. (1989). *Ecology, Community and Lifestyle: Outline of an Ecosophy*. Cambridge: Cambridge University Press.

Næss, A. (1973). The Shallow and the Deep, Long-Range Ecology Movement. *A. Summary. Inquiry* 16 (1-4), 95–100. doi:10.1080/00201747308601682

Nolt, J. (2014). *Environmental Ethics for the Long Term: An Introduction*. New York: Routledge.

Palmer, C. (2003). "An Overview of Environmental Ethics," in *Environmental Ethics: An Nthrology*. Editors A. Light and H. RolstonIII (Malden: Blackwell).

Palmer, C. (2010). *Animal Ethics in Context*. New York: Columbia University Press.

Paro Robots (2021). PARO Therapeutic Robot. Available at: http://www.parorobots.com (Accessed July 15, 2021).

Reeves, B., and Nass, C. I. (1996). *The media Equation: How People Treat Computers, Television, and New media like Real People and Places*. Cambridge: Cambridge University Press.

Regan, T., and Singer, P. (1976). *Animal Rights and Human Obligations*. Englewood cliffs: Prentice-Hall.

Regan, T. (2004). *The Case for Animal Rights*. Oakland: Univ of California Press.

Rottman, J., Crimston, C. R., and Syropoulos, S. (2021). Tree-Huggers Versus Human-Lovers: Anthropomorphism and Dehumanization Predict Valuing Nature over Outgroups. *Cogn. Sci.* 45 (4), e12967. doi:10.1111/cogs.12967

Samuelson, P. A. (1948). Consumption Theory in Terms of Revealed Preference. *Economica* 15 (60), 243–253. doi:10.2307/2549561

Sætra, H. S. (2021a). Confounding Complexity of Machine Action. *Int. J. Technoethics* 12 (1), 87–100. doi:10.4018/IJT.20210101.oa1

Sætra, H. S., and Fosch-Villaronga, E. (2021). Research in AI Has Implications for Society: How Do We Respond?. *Morals & Machines* 1 (1), 60–73. doi:10.3390/healthcare9081007

Sætra, H. S. (2019). "Man and His Fellow Machines: An Exploration of the Elusive Boundary between Man and Other Beings," in *Discussing Borders, Escaping Traps: Transdisciplinary and Transspatial Approaches*. Editors F. Orban and E. Strand Larsen (Münster: Waxman).

Sætra, H. S. (2021b). Robotomorphy: Becoming Our Creations. *AI and Ethics*. doi:10.1007/s43681-021-00092-x

Sætra, H. S. (2020). The Parasitic Nature of Social AI: Sharing Minds with the Mindless. *Integr. Psych. Behav.* 54, 308–326. doi:10.1007/s12124-020-09523-6

Sætra, H. S. (2014). The State of No Nature: Thomas Hobbes and the Natural World. *Ecol. Saf.* 8, 177–193.

Seed, J. (1988). "Beyond Anthropocentrism," in *Thinking like a Mountain: Towards a council of All Beings*. Editors J. Seed, J. Macy, P. Fleming, and A. Næss (Philadelphia: New Society Publishers), 35–40.

Stone, C. D. (1972). Should Trees Have Standing--Toward Legal Rights for Natural Objects. *S. Cal. L. Rev.* 45, 450.

Tavani, H. (2018). Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information* 9 (4), 73. doi:10.3390/info9040073

Winfield, A. (2012). *Robotics: A Very Short Introduction*. Oxford: OUP.

Zimmerman, M. E. (1985). The Critique of Natural Rights and the Search for a Non-Anthropocentric Basis for Moral Behavior. *J. Value Inq.* 19 (1), 43–53. doi:10.1007/bf00151415

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 22 September 2021
doi: 10.3389/frobt.2021.715849

The Virtuous Servant Owner—A Paradigm Whose Time has Come (Again)

Mois Navon *

Department of Jewish Philosophy, Bar Ilan University, Ramat Gan, Israel

Social Robots are coming. They are being designed to enter our lives and help in everything from childrearing to elderly care, from household chores to personal therapy, and the list goes on. There is great promise that these machines will further the progress that their predecessors achieved, enhancing our lives and alleviating us of the many tasks with which we would rather not be occupied. But there is a dilemma. On the one hand, these machines are just that, machines. Accordingly, some thinkers propose that we maintain this perspective and relate to Social Robots as "tools". Yet, in treating them as such, it is argued, we deny our own natural empathy, ultimately inculcating vicious as opposed to virtuous dispositions. Many thinkers thus apply Kant's approach to animals—"he who is cruel to animals becomes hard also in his dealings with men"—contending that we must not maltreat robots lest we maltreat humans. On the other hand, because we innately anthropomorphize entities that behave with autonomy and mobility (let alone entities that exhibit beliefs, desires and intentions), we become emotionally entangled with them. Some thinkers actually encourage such relationships. But there are problems here also. For starters, many maintain that it is imprudent to have "empty," unidirectional relationships for we will then fail to appreciate authentic reciprocal relationships. Furthermore, such relationships can lead to our being manipulated, to our shunning of real human interactions as "messy," to our incorrectly allocating resources away from humans, and more. In this article, I review the various positions on this issue and propose an approach that I believe sits in the middle ground between the one extreme of treating Social Robots as mere machines versus the other extreme of accepting Social Robots as having human-like status. I call the approach "The Virtuous Servant Owner" and base it on the virtue ethics of the medieval Jewish philosopher Maimonides.

Keywords: social robots, artificial intelligence, ethics, jewish thought, virtue, slave

Edited by:
Anne Gerdes,
University of Southern Denmark,
Denmark

Reviewed by:
Matthew Studley,
University of the West of England,
United Kingdom
David Gunkel,
Northern Illinois University,
United States

Correspondence:
Mois Navon
mois.navon@divreinavon.com

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

Received: *27 May 2021*
Accepted: *16 August 2021*
Published: *22 September 2021*

Citation:
Navon M (2021) The Virtuous Servant Owner—A Paradigm Whose Time has Come (Again).
Front. Robot. AI 8:715849.
doi: 10.3389/frobt.2021.715849

INTRODUCTION

"Man is by nature a social animal" (*Politics*, 1253a). So noted Aristotle almost 3,000 years ago. Interestingly, while Aristotle did actually conceptualize automatons that might replace the slave labor of his day (ibid., 1253b), he did not envision that humans might interact socially with these automatons. This is because, in addition to living at a time when human slaves were considered animated tools, he never imagined the sophisticated automatons of the twenty-first century—i.e., social robots, which today come in a vast and growing array of configurations (Reeves et al., 2020), many designed to be social companions.[1] Indeed, the social robots of today are not merely functional automatons, they are emotionally engaging humanoids. And even those not designed to be so, nevertheless manage to trigger our empathy, drawing us to relate to them *as if* they too were, by nature, a "social animal."

It is this "as if" (Gerdes, 2016: 276) condition that brings us to one of the most consternating conundrums in the field of roboethics today, what Mark Coeckelbergh calls, "the gap problem" (Coeckelbergh, 2013; Coeckelbergh, 2020c). When we interact with a Social Robot (SR), a "gap" exists between what our reason tells us about the SR (i.e., it is a machine) versus what our experience tells us about the SR (i.e., it is more than a machine). It is this gap that gives rise to the ethical question that is the subject of this essay: How are we to relate *morally* to social robots—like a machine or more than a machine?

Before attempting to address this question, it is important to define specifically the type of SR that is the focus of this investigation. Social robots are currently powered by artificial intelligence, which enables them to "learn" from their experiences, modify their behavior accordingly, and give the appearance of autonomy—the appearance of beliefs, desires and intentions. These features are the hallmarks of consciousness and what make us, in large part, who we are. But today, the artificial intelligence powering our social robots is entirely artificial—entirely based on mathematics (see, e.g., Domingos, 2018; Boucher, 2019; Brand, 2020: 207; Coeckelbergh, 2020a: 83–94)[2]—the robot only behaves *as if* it has consciousness.

There are hopes, even designs, to make social robots with true human-like second-order consciousness—i.e., to make a sentient, self-aware being that has the capability to think about its own thoughts. However, while this may be the ultimate goal of the AI project, what Ray Kurzweil calls "the singularity," its achievement remains a long way off (see, e.g., Torrance, 2007: 500; Coeckelbergh, 2010a: 210; Wallach and Allen, 2010: 8; Tallis, 2012: 194; Veruggio and Abney, 2012: 349; Prescott, 2017: 5; Sparrow, 2017: 467; Bertolini and Arian, 2020: 45; Birhane and van Dijk, 2020: 210; Hauskeller, 2020: 2). And even the less ambitious HLMI [High-Level Machine Intelligence] is a long way off, see, e.g., Grace et al. (2018), Boucher (2019: 10), and Shalev-Shwartz et al. (2020: 2). Some, however are optimistic: Dyson (2012), Moravec (1988), Kurzweil 1999 cited in Sparrow and Sparrow (2006), Long and Kelley 2010, O'Regan 2012, and Gorbenko et al. 2012 cited in Neely (2013). Accordingly, this paper does not seek to discuss social robots with human-like consciousness, nor even with simple animal sentience,[3] but rather social robots that are driven by current day artificial intelligence—i.e., robots that are essentially autonomous mobile computers with humanlike physical characteristics,[4] what I call: mindless humanoids.

THE DILEMMA

So, again, the question is: How are we to relate *morally* to social robots?

In general, when we encounter a new entity—be it mineral, vegetable, animal, or human—we seek to categorize it according to its various ontological properties (see, e.g., Coeckelbergh, 2013: 63; Johnson and Verdicchio, 2018: 292). We do this so that we know how to interact with it, and more profoundly, how to interact with it morally. For example, if it is a rock, we know we can kick it into an open field without qualms about harming the rock; if it is a neighborhood cat, we know that we shouldn't kick it or otherwise indiscriminately cause it pain; if it is our human co-worker, we realize that greater moral consideration is due him than a cat. In short, we ask what the entity "is" in order to determine how we "ought" to treat it.[5] This approach is variously known as the ontological approach, the properties approach (Tavani, 2018), the mind-morality approach (Gerdes, 2016), the organic approach (Torrance, 2007; Tollon, 2020), the realist approach (Torrance, 2013) or simply, the standard approach (Coeckelbergh, 2013).

The ontological approach, however, encounters difficulties with social robots as they fall into a strange middle ground between man and machine, presenting the previously mentioned gap problem, alternatively referred to as a "category boundary problem" (Coeckelbergh, 2014: 63). On the one hand, the SR is a mindless automaton, programmed[6] to carry out various social tasks—i.e., a machine. On the other hand, the SR, designed with human-like physical characteristics and programmed to carry out its tasks with human-like behavior, appears to us as, well, human-like. Furthermore, even if we are

[1] For the sake of completeness, it should be made clear that Aristotle did envision *intelligent* artificial servants, nevertheless, he could not imagine interacting with them other than as natural slaves, since slaves were a natural part of his politics. His desire for automatons was motivated not by ethical qualms but by expediency (*Politics* 1253b). For more on this see LaGrandeur (2013: 9–11, 106–108).

[2] For the sake of completeness, today's AI is known as Narrow or Weak AI, which uses algorithms to analyze data, mathematically, and make decisions accordingly. This is as opposed to General or Strong AI (sometimes referred to as GAI or AGI), which seeks to make machines intentional with consciousness. How will this be done is of great debate. There are "computationalists" (e.g., Ray Kurzweil, Hans Moravec) who believe that when every brain function is implemented at the level of human brain processing power, consciousness will "emerge." Others (e.g., Pentti Haikonen) explain that it is not just the computational power that is needed but the way the computations are done (e.g., via associative neural networks, etc.). Still others (e.g., Roger Penrose, Colin Hales) believe that computation in itself, in any manner, is not enough but rather the physics of the brain must be replicated for consciousness to emerge.

[3] While there is much to be said in regard to our moral attitude toward sentient robots, such a discussion remains outside the scope of this article.

[4] I make the proviso of "humanlike" to exclude autonomous mobile computers like autonomous vehicles or assembly-line machinery for which I have yet to read of individuals becoming emotionally engaged.

[5] For a concise discussion of the is-ought debate see Gunkel (2018: 3–4). See also Coeckelbergh (2013: 63), Schwitzgebel and Garza (2015: 99).

[6] The term applies whether the SR is driven by conventional programming (i.e., rule based hard-coded algorithms) or machine learning (see, e.g., Domingos, 2018; Boucher, 2019).

aware of the fact that it is not human, that it does not have a mind, a consciousness, we are nevertheless deceived (see, e.g., Turkle, 2011a: 63, 90; Grodzinsky et al., 2014: 92, 98; Richardson, 2015: 124; Gunkel, 2018: 115; Leong and Selinger, 2019: 307).

The deception is of course self-deception, a result of our own human "programming," if you will. We are "wired" to respond to animacy, to self-propelled entities that "make eye contact, track our motion, and gesture in a show of friendship" (Turkle, 2011a: 8; see also, e.g., Arico et al., 2011; Gray and Schein, 2012: 408; Scheutz, 2014b: 213; Darling et al., 2015: 770; Schwitzgebel and Garza, 2015: 112; Darling, 2016: 217; Ghiglino and Wykowska, 2020: 53). These behaviors push, what Sherry Turkle calls, "our Darwinian buttons," inducing us to ascribe human attributes to such robots until we "imagine that the robot is an 'other,' that there is, colloquially speaking, 'somebody home'" (Turkle, 2011a: 8; see also, e.g., Foerst, 2009; Arico et al., 2011; Turkle, 2011b: 63; Scheutz, 2014b: 215; Richardson, 2015: 72; Bertolini, 2018: 649; Fossa, 2018: 124). Sven Nyholm calls this "mind reading"—we read into the behaviors of others their apparent mental state, their mind (Nyholm, 2020; see also, e.g., Richardson, 2015: 74; Darling, 2016: 216; de Graaf and Malle, 2019; Ghiglino and Wykowska, 2020: 51). Others (e.g., Duffy, 2003: 180; Huebner, 2009; Veruggio and Abney, 2012: 355; Ghiglino and Wykowska, 2020: 67; Tollon, 2020: 7) say we adopt, what Daniel Dennett terms, the "intentional stance," whereby we treat an entity "*as if* it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires'" (Dennett, 1996).

This phenomenon of seeing social robots as humanlike is known as anthropomorphism, but it doesn't end with simply ascribing human beliefs, desires and intentions to the robot—we take it to the next step and become engaged, emotionally, with the social robot (see, e.g., Coeckelbergh, 2009; Choi, 2013; Grodzinsky et al., 2014: 92; Darling, 2016: 214; Richards and Smart, 2016: 18; Darling, 2017; Johnson and Verdicchio, 2018; Tavani, 2018: 3; Gunkel and Wales, 2021; see also sources cited in previous paragraph). And this engagement isn't just some kind of fictional role playing, but rather, we feel real empathy toward the social robot (see, e.g., Redstone 2014; Darling et al., 2015; Wales, 2020). Indeed, Tony Prescott notes that "we do not need to believe (or be deceived) that the psychological states, intentional, or phenomenological, that we read into an artefact, such as a robot, are akin to our own in order to experience an authentic and meaningful emotional response" (2017: 144).

Now, while this emotional anthropomorphizing is going on, another socio-psychological element comes into play: dehumanization. Massimiliano Cappuccio et al., describe this troubling phenomenon:

> "... the fundamental ethical problem at the core of social robotics is that, while robots are designed to be like humans, they are also developed to be owned by humans and obey them. The disturbing consequence is that, while social robots become progressively more adaptive and autonomous, they will be perceived more and more as slave-like. In fact, owning and using an intelligent and autonomous agent instrumentally (i.e., as an agent capable to act on the basis of its own decisions to fulfill its own goals) is precisely the definition of slavery" (Cappuccio et al., 2019: 25).

Cappuccio et al. call this the Anthropomorphism Dehumanization Paradox (ADP). Jordan Wales (2020) calls it "the dilemma of empathy and ownership," explaining that if we allow ourselves to engage emotionally with robots, we will nevertheless use them for what we acquired them to do and, accordingly, end up treating them as slaves (similarly, Walker, 2006). This might not seem so terrible since the machine "feels" no indignity or ignominy, no disgrace or denigration—indeed, the machine "feels" nothing.[7] The problem, however, is not for the machine but for man, as Kant famously noted:

> So if a man has his dog shot, because it can no longer earn a living for him, he is by no means in breach of any duty to the dog, since the latter is incapable of judgement,[8] but he thereby damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind. Lest he extinguish such qualities, he must already practise a similar kindliness towards animals; for a person who already displays such cruelty to animals is also no less hardened towards men. We can already know the human heart, even in regard to animals (Kant, 1996, 212).[9]

Similarly, it is feared that our instrumental treatment of human-like robots—treating them as slaves—will then influence our treatment of humans (e.g., Levy, 2009; Anderson, 2011: 294; Darling, 2016: 227–8; Cappuccio et al., 2019: 14; Chomanski, 2019: 1008; Gunkel and Wales, 2021: 4, 9; Coeckelbergh, 2021: 7; in opposition see, e.g., Johnson and Verdicchio, 2018; Bryson, 2020a: 22). We will likely not treat people as slaves, but we will certainly be in danger of treating people as objects rather than subjects. Our relationships with SRs, to put it Buberian terms, could be seen as habituating an I-It relationship as opposed to cultivating an I-Thou relationship (Buber, 1970). The SR would thus invert Buber's call to relate to

[7] The debate on whether it is possible to give machines emotions and feelings is outside the scope of this paper. Suffice it to say that truly sentient machines are not, as mentioned above, in the offing.

[8] Kant famously held that the line dividing those deserving of moral status versus those undeserving of such was "judgement" (or reason), a position which became anathema following Bentham's revision of the dividing line to "sentience," or more precisely, the ability to suffer (Bentham, [1789] 2019). So, while Kant's example of dog may grate on today's sensibilities, it provides a fitting paradigm to address the mindless humanoid which has neither judgement nor sentience.

[9] Worthy of note is that Kant (1724–1804), here, was preceded by Nachmanides (1194–1270) who explains that the biblical command to send the mother bird away before taking her eggs was promulgated in order "that we should not have a cruel heart and lack compassion ... and is to prevent us from acting cruelly" (Nachmanides, 1976: Deut. 22.6). Thus, while some argue that Kant's words point only to a concern for causal action and not character disposition (see fn. 10 herein), Nachmanides explicitly voices concern for both aspects, reiterating, "the reason for the prohibition is to teach us the trait of compassion and that we should not be cruel ... " (ibid.).

the other as subject not object, hardening us, to echo Kant, to view the other as object not subject (Hawley, 2019, 12). And this, ultimately, reflects upon the individual as vicious as opposed to virtuous.[10] For Buber, the individual—the "I"—is not merely influenced by his relationship with the other, he is *defined* by it. "There is no I as such but only the I of the basic word I-Thou and the I of the basic word I-It. When a man says I, he means one or the other" (Buber, 1970: 54). Consequently, some, like Michael Burdett (2020), have suggested that it would be appropriate for us to relate to a robot as a "Thou." Others, like Elizabeth Green (2018) argue that a robot can never be a Thou, while still others, like Sherry Turkle (2011a: 85), explain that the "Thou" relationship simply emerges.

RESOLUTIONS

This brings us into the thick of possible "resolutions" to the dilemma. I keep the term "resolution" in quotes because this dilemma, like all worthy of the name, only reach resolution with the sacrifice of ideals. This point will be made all too clear in the following review of proposed resolutions.

Returning to Cappuccio et al. (2019: 26), who describe the dilemma as a paradox, we encounter two practical approaches to dissolve the paradox: either reduce—by design—the elements that promote anthropomorphizing, thus keeping the machine very much a machine,[11] or conversely, increase those elements that engender empathy to encourage human to human-like interaction.[12] Both approaches, they note, are not really solutions. Reducing the anthropomorphic elements of SRs undermines their very purpose as companions that are to "establish trust and cooperation, [be it] with a child, a patient with disabilities, or an elderly person" (Cappuccio et al., 2019: 26). On the other hand, increasing such elements that engender human-like empathic relationships, opens a Pandora's box of ethical issues based on the misperception of the true nature of the machines, including but not limited to: developing intimate relationships with robots (Turkle, 2011a: 295; Richardson, 2015: 12; Gerdes, 2016: 277; Bertolini, 2018: 653), shunning human relationships as "messy" (Turkle, 2011a: 7; similarly, Whitby, 2008: 331; Bryson, 2010: 7; Toivakainen, 2015: 10), prioritizing humanoids over humans, thus misspending or misallocating resources (Torrance, 2007: 498; Bryson, 2010: 3; Neely, 2013; Schwitzgebel and Garza, 2015: 114), sacrificing human life (Torrance, 2007: 508; Smids, 2020: 2850), seeing oneself as a machine and thus shirking moral responsibility (Metzler, 2007: 20), and generally maintaining a warped view of reality (Sparrow and Sparrow, 2006: 155; Gerdes, 2016: 276).

The two solutions that Cappuccio et al. float can be seen as an attempt to sway a resolution to the gap problem. That is, either we emphasize what our reason tells us about the SR (i.e., it is a machine) or we emphasize what our experience tells us about the SR (i.e., it is more than a machine). Interestingly, this dichotomy reflects the split of the philosophical community in to two distinct camps.[13] On the one side, there is the "instrumental" camp, populated by those who believe that machines are machines and, regardless of their appearance and behavior, we should relate to robots like we would to a toaster or a vacuum cleaner (see, e.g., Gunkel, 2018: Ch. 2 "!S1 !S2"). On the other side, there is the "appearances" camp, populated by those who maintain that it is precisely through appearance and behavior that we engage with others and must similarly relate to robots (see, e.g., Gunkel, 2018: Ch. 5 "!S1 S2").

The instrumental camp could also be referred to as the "insides count" camp, in that they take the position referred to earlier as the "ontological approach." They derive the moral status of the entity based on its ontology, on "what's going on inside." Accordingly, sentience or first-order consciousness is needed for moral patiency and second-order consciousness is needed for moral agency (see, e.g., Anderson, 2013; Smids, 2020). In opposition, the "appearances" camp argues that we have no method to reveal the insides of an entity for we have no "privileged access" to determine if a being is conscious. As a result, we must content ourselves with externals, with the behavior of the entity and its interaction with us. Some here argue that this approach is not simply an accommodation due to epistemological deficiencies but is the philosophically preferred approach based on our lived experience of SRs (see, e.g., Gunkel, 2018; Coeckelbergh, 2010a). Accordingly, we must grant SRs, if not full moral agency then, moral patiency or moral consideration. This approach has been called the relational approach (Coeckelbergh, 2010a; Richardson, 2015) the phenomenological approach (Coeckelbergh 2010b), the hermeneutic approach (Coeckelbergh, 2021), and includes the ethical behaviorist approach (Neely, 2013; Danaher, 2019).

THE MIDDLE CAMP

Now, while I have described the dilemma as being approached from two sides, two camps, there is in fact a middle ground, a middle camp, occupied by thinkers that believe insides count but also believe that there are reasons to relate morally to the mindless humanoid as more than a mere machine. That is, though the SR is

[10] Worthy of note is the disagreement over whether Kant is concerned only with the externally causal effect—e.g., kicking a dog will bring one to kick a human (see, e.g., Coeckelbergh, 2020b; Coeckelbergh, 2020c; Sparrow 2020)—or does Kant's demand for virtuous behavior because it reflects on the character of the individual (see, e.g., Gerdes, 2016; Denis, 2000).

[11] Many make this argument, e.g., Bryson (2010: 65), John McCarthy and Marvin Minsky in Metzler (2007: 15), Miller (2010), Grodzinsky et al. (2014), Schwitzgebel and Garza (2015: 113), Richards and Smart (2016: 21) and Leong and Selinger (2019). The position is even offered as a regulatory principle (Boden et al., 2010: #4), though Wales (Gunkel and Wales, 2021: 11) argues it will simply not be followed.

[12] Many make this argument, e.g., Breazeal (2002), Duffy (2003), Walker (2006), Darling (2017), and Burdett (2020).

[13] Cappuccio et al. (2019: 10) note the two camps explicitly; so too, Torrance (2013: 10). Gunkel (2017, 2018) adds two additional camps in order to account for sentient machines (which, as mentioned, are beyond the scope herein). It should be noted that Gunkel defines yet another camp for himself.

neither a moral agent nor a moral patient, there are nevertheless ethical demands incumbent upon humans in their interactions with it. Steve Torrance, who I place in this middle camp, describes the moral relationship with a robot as "quasi-moral" (2007: 504, 516). I understand this to mean that the moral demands engendered in the HRR (Human Robot Relationship) do not stem from the inherent moral *status* of the robot but from the relationship, from the moral *implications* of the relationship. This, it should be noted, is in contradistinction to the "relational approach" which sees the mindless humanoid as a "quasi-other." To be clear, in the "quasi-other" approach it is otherness, alterity, that is imposed on the robot itself which consequently engenders a very real moral demand—e.g., the demand to treat the other like yourself;[14] whereas in the "quasi-moral" approach, it is morality (e.g., a norm) that is imposed on an otherwise amoral situation.

This quasi-moral approach taken by the middle camp finds its ground in Kant's indirect duties to the animal kingdom. Kant believed that animals have no moral status and accordingly, he writes, "we have no immediate [i.e., direct] duties to animals; our duties towards them are indirect duties to humanity" (Kant, 1996: 212). Anne Gerdes (2016) explains Kant as teaching that we have not duties *to* animals but rather we have duties *with regard to* animals; similarly, reasons Gerdes (as does Bryson, 2010), we have not duties *to* robots but rather we have duties *with regard to* robots. She brings Kant's writing on this point in his *Metaphysics of Morals*:

> . . . a propensity to wanton destruction of what is beautiful in inanimate nature . . . is opposed to a human being's duty to himself; for it weakens and uproots that feeling in him, which, though not of itself moral, is still a disposition of sensibility that greatly promotes morality or at least prepares the way for it. . .
>
> With regard to the animate but non-rational part of creation, violent and cruel treatment of animals is far more intimately opposed to a human being's duty to himself, and he has a duty to refrain from this; for it dulls this shared feelings of their suffering and so weakens and gradually uproots a natural predisposition that is very serviceable to morality in one's relations with other men. . . .
>
> Even gratitude for the long service of a horse or dog belongs indirectly to a human being's duty with regard to these animals; considered as a direct duty, however, it is always only a duty of the human being to himself (6:443).

This passage, as well as the one quoted immediately prior, can be seen as advancing a virtue ethics approach toward non-human entities—as, indeed, Gerdes writes. That is, in our actions toward the inanimate, though no deontological demands bind our behavior, we are nevertheless to refrain from wanton destruction as part our efforts at developing a disposition that promotes moral behavior—i.e., in order to develop our virtuous character (so too, Toivakainen, 2015: 278). With regards to animals, our behavior has an even greater impact on our dispositions. Lara Denis explains that, for Kant, "Any way of treating an animal that could impair our ability to feel love and sympathy for others constitutes a risk to a morally valuable aspect of our rational nature. Kant thinks that cruel or even unloving treatment of animals threatens to impair us in this way" (Denis, 2000: 409). Denis explains that the reason our interactions with animals so affect our dispositions is because we share our animal nature with them and because they engage us emotionally.

Given this, I would argue that, while a SR could be considered an inanimate object, its human-like interaction with us, to the point of our attributing mental states to it, places the SR more closely in the animate category. And though we don't share our biological animal nature with the robot, we do share behaviors engendered by our animal nature (see, e.g., Turkle, 2011a: Ch. 7). Furthermore, while our emotional engagement with the robot lacks the authentic sentient elements of pain and pleasure characteristic of animal interaction, behaviorally we are just as engaged (see prior sources on emotional engagement as well as, e.g., ibid.; Cappuccio et al., 2019: 15–16). Accordingly, without arguing for the "appearances" approach, I am calling for a virtue approach—i.e., an approach which acknowledges and accounts for how the interaction with a mindless humanoid affects the virtue of the human interlocuter.

The virtue approach to robots is not new and has, in fact, been promoted by numerous thinkers such as: Anne Gerdes (2016), Robert Sparrow (2017, 2020), Shannon Vallor (2018), Massimiliano Cappuccio et al. (2019), and even Mark Coeckelbergh (2020b, 2020c, though he argues against in 2010a). However, while virtue ethics clearly eliminates the "dehumanizing" part of the "anthropomorphizing while dehumanizing paradox," it would appear to utterly capitulate to the anthropomorphizing part. That is, by relating to the SR in a virtuous manner we avoid the evils inherent in dehumanizing it but remain susceptible to the previously mentioned Pandora's box of negative consequences associated with anthropomorphizing it. Consequently, Cappuccio et al. (2019: 26) acknowledge that they are thus at a loss to resolve the paradox and content themselves to apply virtue ethics to avoid dehumanizing.

One scholar who does attempt a resolution is Jordan Wales (2020), who employs the thought of Augustine to address the paradox. Augustine, in his *De doctrina Christiana* (1:33:37), teaches that one should ever seek to refer his joy in an other toward God, toward the creator of that individual.[15] Wales applies this notion to our interactions with SRs, such that,

[14] This approach is found in numerous authors, as, for example, the following list shows. Coeckelbergh (2010b): a robot is "quasi-alterity" to be treated as it appears to us. Burdett (2020): a robot is "quasi-person" which demands "Thou" relations. Don Ihde (1990: 100): a robot is "quasi-other" but remains lower than human or animal; see also Bergen and Verbeek (2020). Peter Asaro (2006): a robot is "quasi-moral agent" giving it some level of responsibility. Philip Brey (2014) argues that the term "quasi-moral agent" denotes involvement in moral acts but without true moral responsibility. Gunkel (2018: Ch. 6) argues for Levinasian alterity relations—i.e., a robot is a full other, not simply a quasi-other.

[15] This is a well-known religious technique wherein one is to channel one's emotions toward God in an effort to connect to the source of all emotion and life itself (see, e.g., Horowitz, 1873: Gen. 46:29).

upon feeling natural empathy toward a SR, "we *redirect* that empathy, 'refer' it, as Augustine would say, to all the unknown concrete persons whose interactions have unwittingly sculpted the persuasive personality of this instrument" (Wales, 2020: 7). Wales thus solves the anthropomorphism problem, or more precisely, the empathy problem inherent in anthropomorphizing.

To be clear, in anthropomorphizing mindless humanoids, we are in danger of becoming emotionally engaged with entities that do not warrant such engagement and which can thus lead to many social ills (as noted above). By redirecting the empathy in our emotional engagement with the SR toward the real flesh and blood people who served to create it, Wales argues that we avoid attributing humanity to the robot, allowing our emotions to find their proper terminus in true humanity.[16] As a result, we can interact with the SR in a virtuous way, allowing our natural empathy and anthropomorphizing to occur and yet maintain the realization that the robot is not human, does not have the moral status of a human and does not enter the moral circle of humanity.

Now, while this idea of "referring" or "redirecting" one's intentions is an accepted notion as a religious ideal, allowing for an adherent to utilize an emotional encounter as a means to develop a connection with his creator, it does not, in my humble opinion, work in other contexts. Indeed, even in the religious context, such channeling of thoughts and emotions is not simple and accomplished only by the truly devout (see, e.g., Maimonides, 1956: III:51; Horowitz, 1873: Gen. 46:29). To expect people to "reference" an other through a SR while in the midst of their everyday mundane lives is utterly impractical. To help us envision the idea, Wales analogizes the connection of 'robot-creator(s)–to–robot' to that of 'baker-to-cookie'—i.e., we could "reference" the baker when we eat his cookie. It is certainly nice to contemplate such a notion, but again, utterly impractical. Furthermore, I think a better analogy of 'robot-creator(s)–to–robot', instead of 'baker-to-cookie', would be 'parent(s)-to-child'. This analogy, I believe, makes clear just how terribly difficult it is to redirect or refer one's thoughts to an other—for, can one really focus on the parent(s) of a child while interacting with the child alone—whether upon first thought or, as Wales suggests, upon second thought.[17] Again, as a religious ideal, reflecting upon the creator in an encounter with an other may be a worthy challenge, but to import the technique to robot encounters will simply not work.[18]

An opposing attempt to resolve our dilemma is brought by Raffaele Rodogno (2016). That is, if Wales attempted to solve the dilemma by framing the HRR as very real, the solution offered by Rodogno is to cast it as utterly fictional:

> . . . we could hypothesize that, when engaging affectively with robot pets, individuals adopt a cognitive mode akin to that which is normally adopted in our engagement with fiction. Being emotionally engaged by robot pets would be akin to being emotionally engaged by a good novel or movie. Just as my sadness for Anna Karenina involves my *imagining*, *accepting*, *mentally representing* or *entertaining the thought*, *without believing*, that certain unfortunate events have occurred to her, my joy at the robot pet involves my imagining, accepting, mentally representing or entertaining the thought, without believing, that it is happy to see me (Rodogno, 2016: 11).

This solution is untenable for a number of reasons. First of all, the relationships we build with fictional characters on the page or screen are both temporary and passive—our interaction with them is limited in time and confined in "space" to our own mind. Robot interactions, in contradistinction, are ongoing active relationships with entities deceivingly alive in the three dimensional space in which we live. As such, they are very different not only from fictional storybook characters but even from real dolls that are not animated to the point that we ascribe to them beliefs, desires and intentions (see, e.g., Turkle, 2011a: 39). Secondly, as noted above (sec. 2 The Dilemma), we take these relationships quite seriously, treating them as if they were not merely fictional—a fact that has dangerous consequences, as Gerdes notes: "the relational *as if* approach is challenged by the fact that, over time, our human-human relations may be obscured by human-robot interactions" (Gerdes, 2016: 276).

In psychological terms, the HRR engenders a state of cognitive dissonance (Festinger, 1957) wherein one knows he is interacting with a very real entity, a SR, while at the same time knowing very well that the interaction is not "real," not authentic. Both Wales and Rodogno attempt to diffuse the dissonance, but from opposite ends. Wales attempts to achieve cognitive harmony by relating the relationship to something real, authentic. That is, since the physical interaction is real, he tries to make the metaphysical relationship real as well. It doesn't work because the referred metaphysical relationship can't be imagined. Attacking the problem from the other end, Rodogno attempts to achieve

[16] Burdett (2020: 355), basing himself on Pattison, makes a similar point. All of these thinkers have been preceded, in a sense, by Buber (1970: 175) who, upon confronting a Doric column in a Syracuse church, writes that he related to the "spiritual form there that had passed through the mind and hand of man and become incarnate." A distinction worthy of note is as follows. Buber is seeking to establish the I-Thou relationship with the inanimate by "referring" to the humanity behind it—he is trying to generate a close, "Thou", relationship; while Wales is trying to "refer" the already close "Thou" relationship to its underlying humanity to avoid seeing the robot as more than it is and falling into the misplaced-empathy trap.

[17] Wales attempts to make the creators of the robot more resident in the robot by explaining that it is not the engineers who built the robot that are represented in the robot, but the very people whose behaviors made up the data that was used to train the neural network that grounds the robot's behaviors. However, the same could be said of the child whose behaviors are made by the DNA and parental education that make up the neural network that grounds the child's behavior. In any case, the notion of referencing is not practical.

[18] I make this claim as a religious man who appreciates the religious ideal. I am not alone in this claim, for when I made it directly to Wales at the RP2020 conference (as he notes in his fn. 22), many other voices joined me in dissent and none his in defense.

cognitive harmony by framing the relationship as completely fictional, inauthentic. That is, since the metaphysical relationship is fictional, he tries to make the physical relationship fictional as well. It doesn't work because the physical relationship can't be imagined away.

VIRTUOUS SERVANT OWNER

And so we return to our question: How are we to relate *morally* to social robots?

Having reviewed the various attempts to construct a response, it is clear that the question, in both physical and metaphysical terms, is strained in the tension between the need to preserve virtue, on the one hand, and the need to preserve authenticity, on the other—what might be termed the Virtue-Authenticity Dialectic (VAD). The ideal response, then, must strive to allow us to maintain our virtuous character, such that we not act in dehumanizing ways toward SRs, but at the same time allow us to maintain our appreciation for authenticity, such that we not accustom ourselves to "as if" relationships *as if* they were real.

As for the "virtue" part of the response, Aristotle's virtue ethics, as echoed in Kant's appeal to indirect duties toward animals, soundly satisfies this need as evidenced by its broad support among thinkers in the field. As for the "authenticity" part of the response, thinkers in the field, as noted, run into trouble.

To address the "authenticity" issue, it is instructive to revisit Aristotle's approach to automata as found in his *Politics*:

> Now of instruments some are inanimate and others animate—the pilot's rudder, for example, is an inanimate instrument, but his lookout an animate one; for the subordinate is a kind of instrument whatever the art . . . if each of the instruments were able to perform its function on command or by anticipation, as they assert those of Daedalus did, or the tripods of Hephaestus (which the poet says "of their own accord came to the gods' gathering"), so that shuttles would weave themselves and picks play the lyre, master craftsmen would no longer have a need for subordinates, or masters for slaves (Aristotle, 2013: 1253b).

Aristotle here envisions that automata will replace slaves as instruments of their masters (similarly, *Nichomachean Ethics* 1161b). Now, while Aristotle may have been the first to articulate this instrumental approach, the history of automata, real or fictional, leaves little doubt that automata were forever imagined to be slaves (see, e.g., LaGrandeur, 2013). And with the advent of AI they continue to be so imagined. Hans Moravec claimed, 'By design, machines are our obedient and able slaves' (Moravec, 1988: 100); Nick Bostrom argued that "investors would find it most profitable to create workers who would be 'voluntary slaves'" (Bostrom, 2014: 167); but no one popularized the notion more than Joanna Bryson (2010) who entitled her article on the issue, "Robots Should Be Slaves." Her claim received no small amount of pushback given the cultural scars left on society by the brutal history of human slavery (Bryson, 2020b).

And that brings us to the heart of the matter, for while it is clear that the goal of automation is to relieve humans of their burdens,[19] slavery is an institution that runs counter to modern values. Slavery is an institution that, despite Aristotle's justifications (*Politics*, Book 1, Chs. 4–5), has been shown to undermine the very virtue ethics that Aristotle sought to foster. Powerful evidence of this can be seen in the testimony of Fredrick Douglass (1845) who wrote of his experience as a slave under a woman he refers to here as "my mistress"—i.e., "female master" slaveholder:

> My mistress was, as I have said, a kind and tender-hearted woman; and in the simplicity of her soul she commenced, when I first went to live with her, to treat me as she supposed one human being ought to treat another. In entering upon the duties of a slaveholder, that [now] I sustained to her the relation of a mere chattel, and that for her to treat me as a human being was not only wrong, but dangerously so. Slavery proved as injurious to her as it did to me. When I went there, she was a pious, warm, and tender-hearted woman. There was no sorrow or suffering for which she had not a tear. She had bread for the hungry, clothes for the naked, and comfort for every mourner that came within her reach. Slavery soon proved its ability to divest her of these heavenly qualities. Under its influence, the tender heart became stone, and the lamblike disposition gave way to one of tiger-like fierceness (1845: 32).[20]

Accordingly, as described previously, many have expressed concern that modern robots designed to serve humans will be treated as slaves and engender a moral calamity for their owners.

But is this outcome not unavoidable? Kant believed it is. He wrote that while one must not hold a slave because, in so doing, one violates the freedom that is at the essence of the individual as a person, nevertheless, one could come to an agreement into which the servant enters of his own freewill and can exit of his own freewill. In such a case, Kant, in his *Metaphysics of Morals*, writes:

> Servants are included in what belongs to the head of a household, and, as far as the form (the way of his being

[19] There is a vast literature on how automation, and specifically AI, will replace human labor, see, e.g., LaGrandeur (2013: 161), Marr (2017), Harari (2019: Ch. 2), and Coeckelbergh (2020a: 136).

[20] Similarly, this slave girl testimony: "I can testify, from my own experience and observation, that slavery is a curse to the whites as well as to the blacks. It makes the white fathers cruel and sensual; the sons violent and licentious; it contaminates the daughters, and makes the wives wretched" (Jacobs, 2020); as well as that of French philosopher Alexis de Tocqueville, "Servitude, which debases the slave, impoverishes the master" (de Tocqueville [1835] 2013).

> in possession) is concerned, *they are his by a right that is like a right to a thing*; . . . But as far as the matter is concerned, that is, what use he can make of these members of his household, *he can never behave as if he owned them* (6:284. *Emphasis added*).[21]

Kant here claims that you can maintain a relationship in which, on the one hand, you are in the position of a servant owner; yet, on the other hand, your behavior toward your servant never expresses this position. I believe that we can reconcile Kant's claim with the seemingly damning evidence brought by Douglass to the contrary, as follows.

Douglass wrote: "In entering upon the duties of a slaveholder, she did not seem to perceive that [now] I sustained to her the relation of a mere chattel, and that for her to treat me as a human being was not only wrong, but dangerously so. Slavery proved as injurious to her as it did to me." That is, only upon fully accepting the slaveholder role—in which one relates to the slave as chattel and in which treating a slave as a human being is "not only wrong, but dangerously so"—does slaveholding becomes injurious to the slaveholder. The injury to the slaveholder, then, is when the slaveholder assumes that one must treat the slave as non-human. That is, it was not the owning of a slave per se, but the social concepts of the time that dictated *how* one needed to treat a slave—i.e., by force of "tiger-like" subjugation to ensure obedience.

A machine programmed for obedience, however, would never occasion its owner to impose her will. Nevertheless, there remains a further moral concern in owning a slave, humanoid or human:

> There is some harm to one's own higher moral values and moral character if one establishes oneself as master. . . The problem of using and treating machines as slaves is that one perpetuates a value that sustains the inappropriate agent character, seeing the world and its denizens as one's slaves. You simply should not treat the world as a place in which your will is absolute. You thereby only strengthen that absolutist, disregarding will (Miller, 2017: 5; similarly Coeckelbergh, 2021: 7).

This harkens back to Kant's dog and the concern against habituating vicious character through vicious behavior. In employing machine-slaves, as stated at the outset: we will likely not treat people as slaves, but we will certainly be in danger of treating people as objects rather than subjects. Accordingly, Kant is not concerned for the virtue (or loss thereof) of one who maintains a servant, as long as she behaves toward her servant as a human being and not as "a thing." Sven Nyholm writes that "Kant himself thought that having a human servant does not need to offend against his formula of humanity [i.e., that one must treat others as ends and not merely as means]—so long as the servants are treated well and with dignity" (2020:192).

This idea finds precedence in the legal writings of the Medieval Jewish philosopher Moses Maimonides. He not only preceded Kant in demanding that servants be treated with dignity, he also elaborated such treatment with details that are instructive in both pragmatic and moral dimensions. Here is his original text (*Laws of Slaves* 8:9), interleaved with some clarifications of mine:[22]

> *It is permissible to work a heathen slave relentlessly.* [Biblical law often promulgates rules in concert with ancient custom while nevertheless seeking to provide a moral improvement on the accepted state of affairs (see, e.g., Korn, 2002; Rabinovitch, 2003; Lamm, 2007; on slavery see, e.g., Shmalo, 2012). As such, the strict letter of law allows for slavery but with various moral restraints.[23] The law, however, is seen as a starting point, a floor and not a ceiling, to use the words of Rabbi J. D. Soloveitchik. Accordingly, Maimonides starts with the legal "floor" only to show that we should—and must—rise far above it. It is interesting to note that Kant (*Metaphysics* 6:284) used the same format, starting with the letter of the law allowing for ownership only to then argue for virtue].
>
> *Though this is the law, the quality of virtue and the ways of wisdom demand of a human being to be compassionate and pursue justice, and not make heavy his yoke on his slave nor distress him.* [Maimonides, here, raises us off the floor of the law, outlining his thesis that calls for virtue and justice. He will now elaborate on these two categories, bringing proof texts to support his claims].
>
> *He should give him to eat and drink of every food and drink. The sages of old had the practice of sharing with the slave every dish they ate. And they would provide food for their animals and slaves before partaking of their own meals. As it is said, "As the eyes of slaves follow their master's hand, as the eyes of a slave-girl follow the hand of her mistress, [so our eyes are toward the Lord our God, awaiting His favor]."* [Here Maimonides provides concrete actions toward maintaining virtuous

[21] An important aside: Kant's contract binds the servant but nevertheless allows him to quit. The servant is then like a slave in the sense that he is the property of, and at the command of, the owner, all the while retaining some human dignity in his ability to exercise his will to both enter and exit the contract freely. In reality, however, it would seem that someone in a position to accept such a contract would be in such dire straits that he will likely never have the means to exit the contract. As such, he is only a "free" servant in name but a slave in practice. Furthermore, it is not clear how the owner can unilaterally, according to Kant, "fetch servants back" (ibid.), if the servants are allowed to terminate the contract at will. The only way this makes sense is by saying that the servant failed to give notice when he left. But why would he not give notice and leave legally if he could do so at will? Maybe the giving notice of leave is actually very limited. It seems that Kant's ownership is closer to slavery than would at first appear.

[22] A detailed analysis of this text is being prepared for publication by the author.

[23] For example, killing a slave entails capital punishment (Ex. 21:20, Rashi ad loc.), a slave is set free if injured (Ex. 21:26-27, Kid. 24a), a slave rests on the Sabbath (Ex. 20:9); a runaway slave is not to be returned (Deut. 23:16). On the differences between ancient slavery versus that of the Torah, see Beasley (2019).

interactions, grounded in a verse equating master and slave in their shared neediness].

Nor should a master disgrace his servant, neither physically nor verbally; the biblical law gave them to servitude, not to disgrace. And one should not treat him with constant screaming and anger, but rather speak with him calmly and listen to his complaints.[24] [Clearly the servant is not to be treated merely as a means but as an end. (I wonder if even Kant would have made such a list of directives to regulate the owner).] *This is explicitly stated with regard to the positive paths of Job for which he was praised: "Have I ever shunned justice for my servants, man or maid, when they quarreled with me… Did not He who made me in my mother's belly make him? Did not One form us both in the womb?" (Job 31:13,15).* [The claim here is for just relations, supported by the verse that notes the physiological identity of master and slave].

Cruelty and effrontery are not frequent except with the heathen who worship idols. The progeny of our father Abraham, however, the people of Israel upon whom God bestowed the goodness of the law (Bible), commanding them to observe "righteous statutes and judgments" (Deut. 4:8), are compassionate to all. [Maimonides defuses any claims that come to justify slavery merely because such treatment is "accepted practice" among the nations of the world. This is not some parochial diatribe against non-Jews,[25] but rather part and parcel of his argument for just relations with one's servant, here made irrespective of the inherent value of the servant. That is, justice is incumbent upon the master for the sake of his own virtue and character].

Accordingly, regarding the divine attributes, which He has commanded us to imitate, the psalmist says: "His tender mercies are over all His works" (Psalms 145:9). [Here, as part of his thesis that one must move beyond the strict letter of the law in the treatment of one's servant, Maimonides reminds us of the ethical imperative to strive to imitate the divine virtues, chief among them being that of mercy/compassion. This claim, like the previous one, is incumbent upon the master irrespective of the inherent value of the slave. Worthy of note is that the support verse does not say that God's "mercies are upon all His creatures" but "upon on all His works." Could this not be understood to allow for application to humanoids?]

Whoever is merciful will receive mercy, as it is written: "He will be merciful and compassionate to you and multiply you" (Deut. 13:18). [Maimonides concludes his call for virtue with a religious principle known as "measure for measure," which states that in the measure, or manner, that you act towards others, so too, in the same measure, will God act towards you. Accordingly, even if one does not appreciate the value of a virtuous character, one will certainly appreciate the selfish need of God's mercy. In addition, this call to mercy, to virtue, is made independent of the worth of the servant. It pleads for virtue saying: though you may not recognize the worth of your servant, nor even the worth of your own character, at least recognize your need for mercy and be merciful.]

This text stands as a powerful call to virtue in general, and to virtuous behavior with one's servant in particular. Maimonides here speaks to any and all, regardless of what "stage on life's way" one might be. Indeed, his arguments for virtuous behavior can be seen as addressing the individual in each of the three Kierkegaardian stages of existence, stages in which one is driven by the corresponding motivations: aesthetic, ethical and religious.[26] Starting with the ethical, being that it is the universal—applying to all and in which all struggle (Kierkegaard, 1985: 83), Maimonides enjoins virtue based on the human dignity inherent in the servant as a human being. Moving to the higher motivation of the religious, Maimonides calls for the master to exhibit virtue both because he is a God fearing individual who, like Abraham,[27] accepts the divine ethical norms of the Bible and furthermore, because he is to emulate the attributes of the creator, mercy being primary among them.[28] Maimonides concludes with an appeal to self-interest (i.e., the Kierkegaardian aesthetic), arguing, in essence, that even if one is not moved by these higher motivations, one should act mercifully that he too will be treated mercifully.

Not satisfied in leaving his readers with "mere" motivations, Maimonides takes pains to prescribe practical action. He instructs the master to feed his slave with "every dish" that he himself eats, thus raising the slave to the dignity of the master. He directs the master to feed his slave before he himself sits to eat, thus instilling compassion toward he who is not in charge of his own food. He warns the master to "speak calmly and listen to the slave's complaints," thus changing the very relationship from one of master-slave to one more akin to employer-employee (and a quite considerate employer at that). Maimonides thus transforms ethical ideal into ethical practice which, ultimately, shapes ethical character (Aristotle, [350 BCE] 2004: 23; Ha-Levi, [1523] 1978: Precept 16; Vallor, 2018: 3.3; Cappuccio et al., 2019; Coeckelbergh 2020b).

Of course no ownership, no matter how virtuous, can be justified today. Slavery is an institution that is anathema in modern moral thought and given circumscribed sanction in the bible, due only to ancient cultural mores. Jewish thought has ever sought to ameliorate

[24] Interestingly, in terms of a model for SRs, this would demand that the SR give negative feedback, and as Kate Darling suggests, "respond to mistreatment in a lifelike way" (Darling, 2016: 228; similarly, Cappuccio et al., 2020).

[25] Worthy of note is the great esteem in which Maimonides holds non-Jewish thinkers, frequently quoting, Aristotle and Al Farabi.

[26] Worthy of note is that Maimonides (1956, 3:33) appears to refer to these categories in articulating the "ultimate causes of the Law": 1) the rejection and reduction of the fulfillment of desires—i.e., aesthetic, 2) the promotion of virtuous interaction between men—i.e., ethical, 3) the sanctification of its followers—i.e., religious.

[27] Like Kierkegaard, Maimonides references Abraham as the father of faith; yet unlike Kierkegaard, Maimonides, indeed Judaism in general, does not accept the notion of a religious leap of faith as requiring a teleological suspension of the ethical (see Navon, 2014).

[28] The two demands could be seen to reflect the two levels of the "religious" articulated by Kierkegaard (see Broudy, 1941: 306).

the master-slave relationship (see, e.g., Shmalo, 2012) to the point that Maimonides demands not simply that one treat his servant as an end, but that one treat him as nothing less than a contemporary! He does so, as mentioned, by providing clear practical behaviors underpinned by clear philosophical reasoning, (albeit) based on biblical verse. Significantly, his arguments are not found not in his philosophical writings but in his legal writings, thus giving them normative import and evincing, essentially, a law to go beyond the law.

And this brings us back to SRs. My point here is not to argue for even this most virtuous form of human slavery, but to apply the Maimonidean paradigm—what I call the "Virtuous Servant Owner" (VSO)—to Human Robot Relationships. For, though the virtuous practices demanded by Maimonides address, in part, the biological needs of a human servant (e.g., *feed the servant every dish the master is fed*), the practices, in general, express the need for dignity, compassion and consideration—practices that every virtuous individual must pursue, whether his interlocuter is human or, as is my thesis, humanoid. Accordingly, while *feeding the servant first* is not relevant, saying "please" and "thank you" is relevant, part and parcel of the requirement to *speak calmly*. Similarly, while *feeding the servant every dish the master is fed* is inapplicable, not raising one's voice in anger nor one's hand in violence is most applicable, falling under the rubric of *not disgracing the servant verbally or physically*.

It is my contention that this master-slave relationship delineated by Maimonides provides an eminently reasonable paradigm for interacting with the social robot, one that can provide a resolution to the VAD (as well as the ADP). Starting with the "virtue" part of the "Virtue Authenticity Dialectic", the VSO model demands that we abide by the highest ideals of a virtuous relationship, thus distancing us from the dehumanization trap. This, of course, is the approach taken by Cappuccio et al., and really the whole "appearances" camp, which leads to the problems associated with anthropomorphizing. However, whereas Cappuccio et al., shun the slave-like relationship as "disturbing," VSO embraces it in virtue. VSO defines the SR as our slave, our property, our instrument, all the while commanding us to behave virtuously with it, treating it as an end. Relating to the SR not merely as an instrument, but as an end, allows us to maintain our own virtuous character. Keeping the SR on the level of instrument, allows us to avoid bringing it in to our moral circle and thus avoid *most* of the Pandora's box of misplaced moral status issues.

I say "most" because we are still left with the "authenticity" part of the "Virtue Authenticity Dialectic." That is, if we are interacting with the SR as an end, treating it in the most virtuous of ways, we will, in the words of Turkle, "imagine that the robot is an 'other'"—i.e., a being to engage with emotionally. How, then, can we retain our appreciation for authentic, reciprocal, relationships—relationships in which both parties understand, in the deepest sense, what they themselves are thinking, saying, and doing?[29] How can we remain cognizant of the value of mind-ful humans over mind-less humanoids?

I suggest that it is precisely by framing the relationship in terms of master-slave that we maintain our distance and are ever brought back to the reality that we are interacting with a machine and not the noblest of creations—a conscious human being. The VSO paradigm holds that, while we maintain a virtuous relationship with the SR, we nevertheless bind that relationship in the rubric of master-slave. In so doing, we are forced to abandon the thought that we are having an authentic relationship for the simple reason that such would imply we are, in fact, slaveholders! This would then implicate us as being in violation of the fundamental principles we hold dear: freedom and equality for all humanity. It is, then, the very designation—"slave"—that awakens in us the realization that the relationship with the SR is not authentic, that "insides count" and that authenticity is precious, to be found only in conscious beings.

And is this not what the name robot was supposed to denote from its very beginning? Karl Capek coined the name robot from the Czech word robota meaning "forced labor." But the name robot has since lost its original intent and so a more telling appellation is of the essence. "Slave," though repugnant to modern ears, is really the term that drives home the idea of the robot, for it is precisely this repugnance that allows us to use the SR as the tool it was made for and not as the friend it appears to be.[30] Nevertheless, due to the negatively charged nature of the term (see, e.g., Miller, 2017: 298, Gunkel, 2018: 131), I suggest we use the "less polarizing" term, to quote Gunkel (ibid.: 130), of "servant." And while thinkers such as Coeckelbergh (2015: 224) question if there is a difference in the terms, I believe there is a world of difference—one that turns on Kant's prescription for human relations. Slave implies chattel, treated as a mere means. Servant implies worker, with the potential to be treated as an end (see, e.g., Bryson, 2010). Slave, according to Steve Petersen (2007: 45), implies working against one's will; servant implies *wanting* to work. Certainly a mindless humanoid cannot be considered as working against its will, for it has no "will," and though it similarly has no "wants," by being programmed to serve it could be considered, anthropomorphically, as *wanting* to serve.[31]

GETTING THE METAPHOR RIGHT

That said, whether slave or servant, the metaphor has given rise to numerous objections. Objections that, as Joanna Bryson has contended in her now infamous piece "Robots Should be Slaves," eventuate from failure to "get the metaphor right." By this she refers to the fact that metaphors are imprecise. We use metaphors as tools, conceptual tools, that allows us to think about things we don't know by comparing them to things we do know. But metaphors, by definition, are limited—"there is an apparent claim of identity, but . . . only with respect to certain characteristics" (Ortony, 1975: 52; see also, Jones and Millar, 2017: 604). Accordingly, the slave metaphor is to be used to address the question of the moral interaction with mindless humanoids not as if it entailed identity but only as a rough conceptual paradigm.

[29] On the importance of authentic reciprocal relationships, see, e.g., Turkle (2011a: 6, 2011b: 64), Richardson (2016: 51), Prescott (2017: 143), Bertolini and Arian (2020), and Nyholm (2020: 111–2). Similarly, Veruggio and Abney (2012: 355).

[30] And marking the SR as non-human, or even making it look completely non-human, is untenable because of the great advantages to having them as humanlike as possible (see, e.g., Scheutz 2014b: 209; Ghiglino and Wykowska 2020: 55).

[31] It should be noted that Petersen argues for the moral legitimacy of engineering mind-ful humanoid servants whereas I am merely discussing mind-less humanoids. Elsewhere (Petersen, 2017) he notes that mindless robots certainly have no moral patiency.

And this is where thinkers, as described by David Gunkel, run in to trouble; for, in the effort to demonstrate that robots should not be slaves, that the slave metaphor "may be the wrong metaphor" (2018: 131), the metaphor is assumed to entail identity—i.e., that what is true for human slaves is true for robots. To take but one example, it is explained that slaves have criminal responsibility in Jewish, Roman and United States law, yet applying this to robots is problematic since punishment works only if something matters to the punished (ibid: 123-5). The metaphor is thus stretched to imply its failure. But "getting the metaphor right" means applying it judiciously.

Bryson (2020b) herself writes: "The mistake I made with that title ["Robots Should be Slaves"] was this belief that everyone was sensitive to the truth that you can't own people. The word slave here is about something else." That is, the metaphor only goes so far, robots are to be slaves in the sense that their function is to serve human needs and in the sense that they have no responsibility for their actions and in the sense that we have no direct moral responsibilities toward them (similarly, Grau, 2011: 458).

Veruggio and Abney note that, indeed, it is impossible to apply all of the moral implications latent in the term "slave" to mindless humanoids, for "in reality, our robots are not (for now, anyway) our 'slaves' *in any robust sense*, as they have no will of their own" (Veruggio and Abney, 2012: 352, *emphasis added*). Again, any use of the term "slave" can only be applied in a very limited sense—as found, for example, in computing terminology wherein slaves and masters are simply logic agents, the former accepting and executing commands at the request of the latter.

Veruggio and Abney explain that we view our relationship with robots incorrectly, incoherently, because we are "driven by our collective guilt over the history of slavery" (ibid). Now, while numerous authors have used this guilt driven approach to argue against the slave metaphor (see, e.g., Lavender, 2011; Dihal, 2020) no one has argued the point more obdurately than Gregory Jerome Hampton (2015). Hampton begins by noting that the motivations for robots are the same as for slavery—i.e., cheap labor requiring the "human" touch, one that combines intelligence and dexterity. Though this is true enough, he extrapolates from here to argue that the deployment of robot slaves is identical to the deployment of human slaves. The claim is fallacious because, as Veruggio and Abney noted, robots have not a will of their own.[32] The deployment of mindless humanoids, then, is more like the deployment autonomous cars—the likes of which no one imputes with slavery.

Hampton goes on to express the fear, without providing support, that the deployment of robot slaves will prompt racism. Now, while there is a concern that mistreating robots that impersonate a specific race (or gender) will "confirm and proliferate" such behavior in society at large (Coeckelbergh, 2021: 7), it is hard to see why racism (or misogyny) would emerge otherwise—i.e., without mistreatment or without impersonation. That said, it could be argued that speciesism against robots could emerge, for people do unfortunately harbor ill will toward the other (see, e.g., Gunkel, 2012: 207; Kim and Kim, 2012; Scheutz, 2014a: 249, Musiał, 2017: 1093). But even if speciesism were to result from deploying robot slaves, there is no reason to believe that this speciesism would prompt racism. Peter Singer (2009), who argues that humans exhibit speciesism against animals, does not argue that it has prompted or contributed to racism. He does say that all such prejudices are "aspects of the same phenomenon"—i.e., unjustifiably maintaining oneself as superior over an other (Yancy and Singer, 2015). So one could raise the concern that relating to mindless humanoids as slaves will inculcate a vicious character that could harden us, to echo Kant once again, in our interactions with human beings in general, but not toward one race in particular. But this concern over inculcating a vicious character is one that has already been raised and addressed directly by the VSO paradigm which demands virtuous behavior toward humanoids (as explained in the VSO section above).

Another claim against deploying humanoid robots as slaves is made by Kevin LaGrandeur (2011: 237) who applies Aristotle's warning to beware of powerful slaves who will revolt. That is, once slaves become more powerful than their masters—be they human or humanoid—they will revolt. This may be an issue for "strong AI," as LaGrandeur states, but a mindless humanoid, while more powerful than humans in many respects, does not have an autonomous will to revolt, indeed, does not have an autonomous will period. Accordingly, this concern is of no consequence with respect to mindless humanoids.

That said, LaGrandeur argues that the mere interdependency of slave-systems with their human operators gives rise to what could be considered a "slave revolt" in the sense that the systems are delegated so much control that humans no longer control or even understand what the slave-systems are doing. We are reminded here of Hegel's master-slave dialectic in which masters, by dependence on their slaves, lose touch with reality (Hegel, [1807] 2019). Mark Coeckelbergh, in his "The Tragedy of the Master: Automation, Vulnerability, and Distance" (2015), applies this dialectic to automation in general, and to AI and robots in particular, explaining that robots as slaves will bring upon us the tragedy of which Hegel warned: dependency on automation and alienation from nature. While this may indeed be true, it is neither a reason to stop the advance of automation nor to dissuade use of the master-slave paradigm. For, though the robot as slave, as with all automation, may bring dependency and alienation, it will also provide the boon of freedom from all the burdens inherent in taming nature to human needs. And employing the robot as slave will no more entail these negative "Hegelian" consequences than relating to the robot as companion—in any case, the very automation will engender dependency and alienation. That is the price of freedom from our burdens.

Additionally, Coeckelbergh (2015) argues against using the slave metaphor for we thus limit "the range of human–technology relations" when there are "different roles for, say, robots." While clearly there are many roles robots can play, in speaking about SRs, they all assume human-like roles—whether as care-takers of the elderly, cleaning maids, teachers or hotel concierge—and they all accommodate the servant metaphor without inappropriately reducing the range of relations. The only role that the slave metaphor limits is "companion," and this role, I believe, is one that should be proscribed. For, engaging socially with robo-companions may lead to the social catastrophe of shunning human companions, as Turkle notes, because they are "sometimes

[32] One could argue in his defense that he is, in fact, referring to mindful robots, however he writes explicitly that he refers to "anything resembling an independent consciousness" (2015: x), which readily includes mindless humanoids, as noted in my Introduction.

messy, often frustrating, and always complex" (2011a: 7, 295; see also, e.g., Richardson, 2015: 12; Gerdes, 2016: 277; Bertolini, 2018: 653).

Now, while many of the above arguments against using the slave metaphor are based on the "dehumanizing" nature of the term, Birhane and van Dijk (2020) argue that the metaphor should be eschewed because it "humanizes" the machine. That is, the term "slave," while clearly dehumanizing when applied to mind-ful humans, is paradoxically humanizing when applied to mindless humanoids. By calling a robot a "slave," they claim, we employ a term reserved for humans and thus implicitly make it human; and as a result, we then find ourselves in the immoral position of a slaveowner. To their claims I have two responses. First, the term does not serve to humanize the humanoid any more than our own natural anthropomorphizing of it does—i.e., in any case, as noted above, we "humanize" it. Second, and more to the point, the fact that we will find ourselves in the immoral position of slaveholder is a welcome implication, as explained previously, that forces us to abandon the illusion that we are interacting with a human being, loathe as we are to be found in violation of the freedoms of a conscious being.

One might counter that many (or most) people will not be so loathe. Yet this is precisely what VSO comes to address. VSO is to be seen as a kind of "user instruction manual" requiring the user/owner to relate to their humanoid servant in a virtuous manner. And while a user manual is no guarantee against user abuses, given that VSO requires the master to "*listen to the complaints*" of his servant, VSO concomitantly requires that the humanoid itself be programmed to provide moral feedback/pushback, reminding the master of his duties (similarly, Darling, 2016; Cappuccio et al., 2020). One can imagine an abusive owner screaming epithets while their robo-servant calmly objects with rational feedback. Will this tame the beast? The answer is irrelevant because such an interchange already removes Birhane and van Dijk's objection that the human will become a slave owner. For, a slave, in the face of such abuse, would cower in submission not persist in moral exhortations and refusal to comply. Accordingly, without an obsequious entity to comply, there is no position for an immoral slaveowner to occupy.

This could, however, lead to the master becoming so frustrated that he "kill" his robo-servant. But there can be no "killing" of a mindless machine, only a powering down. Interestingly, it was precisely due to this moral fallacy that Bryson originally applied the slave metaphor. Shocked that people expressed repugnance at the idea of turning off a mindless humanoid, she went on a campaign to decry the notion that a mindless humanoid had moral patiency (Bryson, 2016). When her efforts failed, she decided to employ the slave metaphor to emphasize that we *can* turn humanoids off. She did not mean to imply that we can kill human slaves but only that we must realize that the humanoid robot is built to serve, that they are, in her words: "tools to extend our abilities and increase our efficiency in a way analogous to the way that a large proportion of professional society, historically, used to extend their own abilities with servants" (2010). The servant metaphor, then, was meant to be applied in the sense that mindless humanoids are like servants functionally, i.e., in the operations they perform. It was not meant to humanize nor to imply an identity to human slaves, and though there is admittedly ambiguity here, she meant just the opposite—i.e., the mindless humanoid has not rights nor feelings nor anything human-like that would engender moral patiency. That, she explains, is "getting the metaphor right."

CONCLUSION

In this essay I have taken up the most unpopular position of defending the indefensible: slavery. Of course, I am in no way, shape, or form, advocating human slavery but rather appropriating the paradigm, the metaphor, if you will, in its most virtuous form to guide human interactions with mindless humanoids. I have taken this position, despite the opposition voiced in much of the philosophic community, because I believe that human authenticity, human worth, and human-human relationships are at stake. If we do not appreciate that we are more than "meat-machines" and that our relationships with each other are more than instrumental, we will fail ourselves as human beings and usher in a world of untold moral calamity. It is a category mistake to equate man and machine. The VSO paradigm counters this mistake by maintaining a clear distinction between man and machine, all the while asking man to cultivate virtue in his interaction with machine.

Does this resolve the dilemma inherent in the Virtue-Authenticity Dialectic? As mentioned before, dilemmas are so designated because they have no perfect resolution. I admit that it is problematic to call an entity that appears human-like a "slave," or even, a "servant." I admit that engaging with human-like SRs makes it difficult to disassociate them from real humans. Nevertheless, given the options, I suggest that being a Virtuous Servant Owner allows us to maintain our own virtuous disposition on the one hand, while preserving our appreciation for human authenticity and authentic relationships, on the other.

Accordingly, whereas Cappuccio et al. sought a way to remove the "alienating representations of slavery," I suggest that it is specifically this alienation that is redeeming. It can allow us to define a new ontological category, not human, not animal, but slave/servant—i.e., animated autonomous tool. And we need not fear the reinstitution of human slavery, for with the introduction of robots as animated autonomous tools, we will eliminate any advantage of human slaves—exactly as Aristotle envisioned.[33]

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the Article/Supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

This article is based on my PhD thesis being written at Bar Ilan University under the guidance of Prof. Hanoch Ben Pazi.

[33] Note that even Mark Coeckelbergh (2015: 227) admits this point.

REFERENCES

Anderson, D. L. (2013). "Machine Intentionality, the Moral Status of Machines, and the Composition Problem," in *Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics*. Editor V. C. Müller (Berlin, Heidelberg: Springer), 321–334. doi:10.1007/978-3-642-31674-6

Anderson, S. L. (2011). "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics," in *Machine Ethics*. Editors M. Anderson and S. L. Anderson (New York, NY: Cambridge University Press).

Arico, A., Fiala, B., Goldberg, R. F., and Nichols, S. (2011). The Folk Psychology of Consciousness. *Mind Lang.* 26 (3), 327–352. doi:10.1111/j.1468-0017.2011.01420.x

Aristotle ([350 BCE] 2004). *Nicomachean Ethics*. Translated by Roger Crisp. Cambridge: Cambridge University Press.

Aristotle (2013). *Politics*. Translated by Carnes Lord. 2nd ed. Chicago: University of Chicago Press.

Asaro, P. M. (2006). What Should We Want from a Robot Ethic? *IRIE* 6 (12), 9–16. doi:10.29173/irie134

Beasley, Y. (2019). The Morality of Slavery. Gush Etzion: Yeshivat Har Etzion. Available at: https://www.etzion.org.il/en/tanakh/torah/sefer-shemot/parashat-mishpatim/morality-slavery.

Bentham, J. ([1789] 2019). *An Introduction to the Principles of Morals and Legislation*. Sydney NSW: Wentworth Press.

Bergen, J. P., and Verbeek, P.-P. (2020). To-Do Is to Be: Foucault, Levinas, and Technologically Mediated Subjectivation. *Philos. Technol.* 34, 325–348. doi:10.1007/s13347-019-00390-7

Bertolini, A. (2018). Human-Robot Interaction and Deception. *Osservatorio Del Diritto Civile E Commerciale, Rivista Semestrale* 2 (December), 645–659. doi:10.4478/91898

Bertolini, A., and Arian, S. (2020). "Do Robots Care?" in *Aging Between Participation and Simulation: Ethical Dimensions of Social Assistive Technologies*. Editors J. Haltaufderheide, J. Hovemann, and J. Vollmann (Berlin: De Gruyter), 35–52. doi:10.1515/9783110677485-003

Birhane, A., and van Dijk, J. (2020). "Robot Rights? Let's Talk about Human Welfare Instead," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, February (New York, NY: Association for Computing Machinery), 207–213. doi:10.1145/3375627.3375855

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. 2010. Principles of Robotics. Available at: https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Boucher, P. (2019). "How Artificial Intelligence Works," in *European Parliament Think Tank* (EPRS | European Parliamentary Research Service). Available at: https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)634420.

Brand, L. (2020). "Why Machines That Talk Still Do Not Think, and Why They Might Nevertheless Be Able to Solve Moral Problems," in *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences*. Editors B. P. Göcke and A. R. Von der Putten (Boston: Brill), 203–217.

Breazeal, C. L. (2002). *Designing Sociable Robots*. Cambridge, MA: MIT Press.

Brey, P. (2014). "From Moral Agents to Moral Factors: The Structural Ethics Approach," in *Moral Status of Technical Artefacts*. Editors P. Kroes and P.-P. Verbeek (Berlin: Springer), 125–142. doi:10.1007/978-94-007-7914-3_8

Broudy, H. S. (1941). Kierkegaard's Levels of Existence. *Philos. Phenomenol. Res.* 1 (3), 294–312. doi:10.2307/2102760

Bryson, J. (2010). "Robots Should Be Slaves," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. Editor Y. Wilk (Amsterdam: John Benjamins Publishing Company), Vol. 8, 63–74. doi:10.1075/nlp.8.11bry

Bryson, J. 2016. Robots Are Owned. Owners Are Taxed. Internet Services Cost Information. *Adventures in NI*, June 23, 2016. Available at: https://joanna-bryson.blogspot.com/2016/06/robots-are-owned-owners-are-taxed.html.

Bryson, J. (2020a). The Coexistence of Artificial and Natural Intelligence. *New York: Digital Future Society*, March 2, 2020. Available at: https://digitalfuturesociety.com/interviews/the-coexistence-of-artificial-and-natural-intelligence-interview-with-joanna-bryson/.

Bryson, J. (2020b). "The Artificial Intelligence of the Ethics of Artificial Intelligence," in *The Oxford Handbook of Ethics of AI*. Editors M. D. Dubber, F. Pasquale, and S. Das (New York, NY: Oxford University Press), 3–25. doi:10.1093/oxfordhb/9780190067397.013.1

Buber, M. (1970). *I and Thou*. Translated by Walter Arnold Kaufmann. New York, NY: Charles Scribner's Sons.

Burdett, M. S. (2020). Personhood and Creation in an Age of Robots and AI: Can We Say 'You' to Artifacts? *Zygon* 55 (2), 347–360. doi:10.1111/zygo.12595

Cappuccio, M. L., Peeters, A., and McDonald, W. (2019). Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition. *Philos. Technol.* 33 (1), 9–31. doi:10.1007/s13347-019-0341-y

Cappuccio, M. L., Sandoval, E. B., Mubin, O., Obaid, M., and Velonaki, M. (2020). Can Robots Make Us Better Humans? *Int. J. Soc. Robotics* 13, 7–22. doi:10.1007/s12369-020-00700-6

Choi, C. Q. (2013). Brain Scans Show Humans Feel for Robots. *IEEE Spectrum: Technology, Engineering, and Science News*, April 24, 2013. Available at: https://spectrum.ieee.org/robotics/artificial-intelligence/brain-scans-show-humans-feel-for-robots.

Chomanski, B. (2019). What's Wrong With Designing People to Serve? *Ethic. Theory Moral Pract.* 22 (4), 993–1015. doi:10.1007/s10677-019-10029-3

Coeckelbergh, M. (2009). Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics. *Int. J. Soc. Robotics* 1 (3), 217–221. doi:10.1007/s12369-009-0026-2

Coeckelbergh, M. (2010a). Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. *Int. J. Soc. Robotics* 3 (2), 197–204. doi:10.1007/s12369-010-0075-6

Coeckelbergh, M. (2010b). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12 (3), 209–221. doi:10.1007/s10676-010-9235-5

Coeckelbergh, M. (2013). The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philos. Technol.* 27 (1), 61–77. doi:10.1007/s13347-013-0133-8

Coeckelbergh, M. (2014). "Robotic Appearances and Forms of Life. A Phenomenological-Hermeneutical Approach to the Relation between Robotics and Culture," in *Robotics in Germany and Japan: Philosophical and Technical Perspectives*. Editors M. Funk and B. Irrgang (Frankfurt Am Main: Peter Lang Edition).

Coeckelbergh, M. (2015). The Tragedy of the Master: Automation, Vulnerability, and Distance. *Ethics Inf. Technol.* 17 (3), 219–229. doi:10.1007/s10676-015-9377-6

Coeckelbergh, M. (2020a). *AI Ethics*. Cambridge, MA: MIT Press.

Coeckelbergh, M. (2020b). How to Use Virtue Ethics for Thinking about the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance. *Int. J. Soc. Robotics* 13, 31–40. doi:10.1007/s12369-020-00707-z

Coeckelbergh, M. (2020c). Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, With Implications for Thinking about Animals and Humans. *Minds Mach.* doi:10.1007/s11023-020-09554-3

Coeckelbergh, M. (2021). Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach. *Int. J. Soc. Robotics*. doi:10.1007/s12369-021-00770-0

Danaher, J. (2019). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Sci. Eng. Ethics* 26 (4), 2023–2049. doi:10.1007/s11948-019-00119-x

Darling, K. (2016). "Extending Legal Protection to Social Robots," in *Robot Law*. Editors R. Calo, A. M. Froomkin, and I. Kerr (MA: Edward Elgar), 213–231. doi:10.4337/9781783476732

Darling, K. (2017). "'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy," in *Robot Ethics 2.0*. Editors P. Lin, R. Jenkins, and K. Abney (NY: Oxford University Press), 173–188.

Darling, K., Nandy, P., and Breazeal, C. (2015). "Empathic Concern and the Effect of Stories in Human-Robot Interaction," in Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (IEEE), 770–775. doi:10.1109/roman.2015.7333675

de Graaf, M. M. A., and Malle, B. F. (2019). "People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences," in 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), 239–248. doi:10.1109/HRI.2019.8673308 de Tocqueville, A. ([1835] 2013). *Democracy in America, Part I*. Translated by Henry Reeve. Available at: https://www.gutenberg.org/files/815/815-h/815-h.htm.

Denis, L. (2000). "Kant's Conception of Duties Regarding Animals: Reconstruction and Reconsideration. *Hist. Phil. Q.* 17 (4), 405–423.

Dennett, D. C. (1996). *Kinds of Minds: Toward an Understanding of Consciousness*. New York: Basic Books.

Dihal, K. (2020). "Enslaved Minds: Artificial Intelligence, Slavery, and Revolt," in *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Editors S. Cave, K. Dihal, and S. Dillon (Oxford: Oxford University Press), 189–212.

Domingos, P. (2018). *The Master Algorithm : How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books, a Member of the Perseus Books Group.

Douglass, F. (1845). Narrative of the Life of Frederick Douglass, an American Slave. *Elegant Ebooks*. Available at: http://www.ibiblio.org/ebooks/Douglass/Narrative/Douglass_Narrative.pdf.

Duffy, B. R. (2003). Anthropomorphism and the Social Robot. *Robotics Autonomous Syst.* 42 (3–4), 177–190. doi:10.1016/s0921-8890(02)00374-3

Dyson, G. (2012). *Darwin Among the Machines: The Evolution of Global Intelligence*. New York: Basic Books.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Foerst, A. (2009). Robots and Theology. *EWE* 20 (2), 181–193. https://www.researchgate.net/publication/273886034_Robots_and_Theology.

Fossa, F. (2018). Artificial Moral Agents: Moral Mentors or Sensible Tools? *Ethics Inf. Technol.* 20 (2), 115–126. doi:10.1007/s10676-018-9451-y

Gerdes, A. (2016). The Issue of Moral Consideration in Robot Ethics. *ACM SIGCAS Comput. Soc.* 45 (3), 274–279. doi:10.1145/2874239.2874278

Ghiglino, D., and Wykowska, A. (2020). "When Robots (Pretend to) Think," in *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences*. Editors B. P. Goc̈ke and A. R. Von der Putten (Boston: Brill), 49–74.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *JAIR* 62 (July), 729–754. doi:10.1613/JAIR.1.11222

Grau, C. (2011). "There Is No 'I' in 'Robot," in *Machine Ethics*. Editors M. Anderson and S. L. Anderson (New York, NY: Cambridge University Press).

Gray, K., and Schein, C. (2012). Two Minds vs. Two Philosophies: Mind Perception Defines Morality and Dissolves the Debate between Deontology and Utilitarianism. *Rev. Phil. Psych.* 3 (3), 405–423. doi:10.1007/s13164-012-0112-5

Green, E. E. (2018). Robots and AI: The Challenge to Interdisciplinary Theology. Doctoral Thesis. Toronto (ON): University of St. Michael's College. Available at: https://tspace.library.utoronto.ca/bitstream/1807/93393/1/Green_Erin_E_201811_PhD_thesis.pdf.

Grodzinsky, F. S., Miller, K. W., and Wolf, M. J. (2014). Developing Automated Deceptions and the Impact on Trust. *Philos. Technol.* 28 (1), 91–105. doi:10.1007/s13347-014-0158-7

Gunkel, D. J. (2012). *The Machine Question Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA: The MIT Press.

Gunkel, D. J. (2017). The Other Question: Can and Should Robots Have Rights? *Ethics Inf. Technol.* 20 (2), 87–99. doi:10.1007/s10676-017-9442-4

Gunkel, D. J. (2018). *Robot Rights*. Cambridge, MA: MIT Press.

Gunkel, D. J., and Wales, J. J. (2021). Debate: What Is Personhood in the Age of AI? *AI Soc.* 36, 473–486. doi:10.1007/s00146-020-01129-1

Ha-Levi, A. ([1523] 1978). *Sefer HaHinnuch: The Book of [Mizvah] Education*. Translated by Charles Wengrov. New York: Feldheim.

Hampton, G. J. (2015). *Imagining Slaves and Robots in Literature, Film, and Popular Culture : Reinventing Yesterday's Slave with Tomorrow's Robot*. London: Lexington Books.

Harari, Y. N. (2019). *21 Lessons for the 21st Century*. London: Vintage.

Hauskeller, M. (2020). "What Is it like to Be a Bot? SF and the Morality of Intelligent Machines," in *Minding the Future. Contemporary Issues in Artificial Intelligence*. Editors B. Dainton, W. Slocombe, and A. Tanyi (New York, NY: Springer).

Hawley, S. (2019). Challenges for an Ontology of Artificial Intelligence. *Perspect. Sci. Christian Faith* 71 (2), 83–95.

Hegel, G. W. F. ([1807] 2019). *The Phenomenology of Spirit*. Edited and translated by Terry Pinkard. Cambridge: Cambridge University Press.

Horowitz, I. (1873). *Beer Yitzhak*. Lvov: A. N. Suss. Available at: https://hebrewbooks.org/31492.

Huebner, B. (2009). Commonsense Concepts of Phenomenal Consciousness: Does Anyone Care about Functional Zombies? *Phenomenol. Cogn. Sci.* 9 (1), 133–155. doi:10.1007/s11097-009-9126-6

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Bloomington, Ind.: Indiana University Press.

Jacobs, H. (2020). *Incidents in the Life of a Slave Girl*. S.L.: Modern Library.

Johnson, D. G., and Verdicchio., M. (2018). Why Robots Should Not Be Treated like Animals. *Ethics Inf. Technol.* 20 (4), 291–301. doi:10.1007/s10676-018-9481-5

Jones, M. L., and Millar, J. (2017). "Hacking Metaphors in the Anticipatory Governance of Emerging Technology," in *The Oxford Handbook of Law, Regulation and Technology*. Editors R. Brownsword, E. Scotford, and K. Yeung (Oxford: Oxford University Press). doi:10.1093/oxfordhb/9780199680832.013.34

Kant, I. (1996). *Lectures on Ethics*. Editors P. Heath and J. B. Schneewind (New York: Cambridge University Press).

Kant, I. (2013). *The Metaphysics of Morals*, Editors M. J. Gregor and R. J. Sullivan (New York: Cambridge University Press).

Kierkegaard, S. (1985). *Fear and Trembling*. Harmondsworth: Penguin.

Kim, M.-S., and Kim, E.-J. (2012). Humanoid Robots as "The Cultural Other": Are We Able to Love Our Creations? *AI Soc.* 28 (3), 309–318. doi:10.1007/s00146-012-0397-z

Korn, E. (2002). Legal Floors and Moral Ceilings: A Jewish Understanding of Law and Ethics. *Edah J.* 2 (2). https://library.yctorah.org/files/2016/09/Legal-Floors-and-Moral-Ceilings-A-Jewish-Understanding-Of-Law-and-Ethics.pdf.

LaGrandeur, K. (2011). The Persistent Peril of the Artificial Slave. *Sci. Fiction Stud.* 38 (2), 232–252. doi:10.5621/sciefictstud.38.2.0232

LaGrandeur, K. (2013). *Androids and Intelligent Networks in Early Modern Literature and Culture Artificial Slaves*. New York, NY: Routledge.

Lamm, N. (2007). "Amalek and the Seven Nations: A Case of Law vs. Morality," in *War and Peace in the Jewish Tradition*. Editors L. H. Schiffman and J. B. Wolowelsky (New York: Michael Scharf Publication Trust of the Yeshiva University Press).

Lavender, I. (2011). *Race in American Science Fiction*. Bloomington: Indiana University Press.

Leong, B., and Selinger, E. (2019). "Robot Eyes Wide Shut," in Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, January 29–31, 2019. doi:10.1145/3287560.3287591

Levy, D. (2009). The Ethical Treatment of Artificially Conscious Robots. *Int. J. Soc. Robotics* 1 (3), 209–216. doi:10.1007/s12369-009-0022-6

Maimonides, M. (1956). *The Guide for the Perplexed*. Translated by M. Friedländer. (New York: Dover).

Marr, B. 2017. The 4 Ds of Robotization: Dull, Dirty, Dangerous and Dear. *Forbes*, October 16, 2017. Available at: https://www.forbes.com/sites/bernardmarr/2017/10/16/the-4-ds-of-robotization-dull-dirty-dangerous-and-dear/?sh=79eec3e03e0d.

Metzler, T. (2007). "Viewing Assignment of Moral Status to Service Robots from the Theological Ethics of Paul Tillich: Some Hard Questions," in AAAI Workshop Technical Report WS-07-07 (Menlo Park, California: The AAAI Press), 15–20. https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-004.pdf.

Miller, K. W. (2010). It's Not Nice to Fool Humans. *IT Prof.* 12 (1), 51–52. doi:10.1109/mitp.2010.32

Miller, L. F. (2017). Responsible Research for the Construction of Maximally Humanlike Automata: The Paradox of Unattainable Informed Consent. *Ethics Inf. Technol.* 22 (4), 297–305. doi:10.1007/s10676-017-9427-3

Moravec, H. (1988). *The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Musiał, M. (2017). Designing (Artificial) People to Serve - the Other Side of the Coin. *J. Exp. Theor. Artif. Intell.* 29 (5), 1087–1097. doi:10.1080/0952813x.2017.1309691

Nachmanides, M. (1976). *Ramban (Nachmanides): Commentary on the Torah*. Translated by C. Chavel. Vol. Deuteronomy. New York: Shilo Publishing House.

Navon, M. (2014). The Binding of Isaac. *Hakirah*. 17, 233–256. https://hakirah.org/Vol17Navon.pdf.

Neely, E. L. (2013). Machines and the Moral Community. *Philos. Technol.* 27 (1), 97–111. doi:10.1007/s13347-013-0114-y

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. New York: Rowman & Littlefield Publishing Group.

Ortony, A. (1975). Why Metaphors Are Necessary and Not Just Nice. *Educ. Theor.* 25 (1), 45–53. doi:10.1111/j.1741-5446.1975.tb00666.x

Petersen, S. (2007). The Ethics of Robot Servitude. *J. Exp. Theor. Artif. Intell.* 19 (1), 43–54. doi:10.1080/09528130601116139

Petersen, S. (2017). "Is it Good for Them Too? Ethical Concern for the Sexbots," in *Robot Sex: Social Implications and Ethical*. Editors J. Danaher and N. McArthur (Cambridge, MA: MIT Press), 155–171.

Prescott, T. J. (2017). Robots Are Not Just Tools. *Connect. Sci.* 29 (2), 142–149. doi:10.1080/09540091.2017.1279125

Rabinovitch, N. (2003). The Way of Torah. *Edah J.* 3 (1). https://library.yctorah.org/files/2016/09/The-Way-of-Torah.pdf.

Redstone, J. (2014). "Making Sense of Empathy with Social Robots," in Sociable Robots And The Future of Social Relations: Proceedings of Robo-Philosophy 2014. Editors J. Seibt, M. Nørskov, and R. Hakli (Amsterdam: IOS Press), 171–178.

Reeves, B., Hancock, J., and Liu, X. (2020). Social Robots Are like Real People: First Impressions, Attributes, and Stereotyping of Social Robots. *Technol. Mind Behav.* 1 (1). doi:10.1037/tmb0000018

Richards, N. M., and Smart, W. D. (2016). "How Should the Law Think about Robots?" in *Robot Law*. Editors R. Calo, A. M. Froomkin, and I. Kerr (MA: Edward Elgar), 3–24. doi:10.4337/9781783476732

Richardson, K. (2015). *An Anthropology of Robots and AI Annihilation Anxiety and Machines*. New York, NY: Routledge.

Richardson, K. (2016). Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines. *IEEE Technol. Soc. Mag.* 35 (2), 46–53. doi:10.1109/mts.2016.2554421

Rodogno, R. (2016). "Robots and the Limits of Morality," in *Social Robots: Boundaries, Potential, Challenges*. Editor M. Nørskov (New York: Routledge).

Scheutz, M. (2014a). "Artificial Emotions and Machine Consciousness," in *The Cambridge Handbook of Artificial Intelligence*. Editors K. Frankish and W. M. Ramsey (Cambridge: Cambridge University Press).

Scheutz, M. (2014b). "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots," in *Robot Ethics: The Ethical and Social Implications of Robotics* (London: MIT Press).

Schwitzgebel, E., and Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Stud. Philos.* 39 (1), 98–119. doi:10.1111/misp.12032

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2020). On the Ethics of Building AI in a Responsible Manner. https://arxiv.org/abs/2004.04644.

Shmalo, G. (2012). Orthodox Approaches to Biblical Slavery. *Torah U-Madda J.* New York, 16. Available at: https://www.jstor.org/stable/23596054.

Singer, P. (2009). *Animal Liberation: The Definitive Classic of the Animal Movement*. New York, N.Y.: Harper Collins.

Smids, J. (2020). Danaher's Ethical Behaviourism: An Adequate Guide to Assessing the Moral Status of a Robot? *Sci. Eng. Ethics* 26 (5), 2849–2866. doi:10.1007/s11948-020-00230-4

Sparrow, R. (2017). Robots, Rape, and Representation. *Int. J. Soc. Robot.* 9 (4), 465–477. doi:10.1007/s12369-017-0413-z

Sparrow, R. (2020). Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might it Be Explained? *Int. J. Soc. Robot.* 13 (1), 23–29. doi:10.1007/s12369-020-00631-2

Sparrow, R., and Sparrow, L. (2006). In the Hands of Machines? The Future of Aged Care. *Minds Mach.* 16 (2), 141–161. doi:10.1007/s11023-006-9030-6

Tallis, R. (2012). *Aping Mankind : Neuromania, Darwinitis and the Misrepresentation of Humanity*. Durham: Acumen.

Tavani, H. (2018). Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information* 9 (4), 73. doi:10.3390/info9040073

Toivakainen, N. (2015). Machines and the Face of Ethics. *Ethics Inf. Technol.* 18 (4), 269–282. doi:10.1007/s10676-015-9372-y

Tollon, F. (2020). The Artificial View: Toward a Non-anthropocentric Account of Moral Patiency. *Ethics Inf. Technol.* 23, 147–155. June. doi:10.1007/s10676-020-09540-4

Torrance, S. (2007). Ethics and Consciousness in Artificial Agents. *AI Soc.* 22 (4), 495–521. doi:10.1007/s00146-007-0091-8

Torrance, S. (2013). Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism. *Phil. Technol.* 27 (1), 9–29. doi:10.1007/s13347-013-0136-5

Turkle, S. (2011a). *Alone Together : Why We Expect More Form Technology and Less from Each Other*. New York: Basic Books.

Turkle, S. (2011b). "Authenticity in the Age of Digital Companions," in *Machine Ethics*. Editors M. Anderson and S. L. Anderson (New York, NY: Cambridge University Press).

Vallor, S. (2018). *Technology and the Virtues a Philosophical Guide to a Future Worth Wanting*. New York, NY: Oxford University Press.

Veruggio, G., and Abney, K. (2012). "Roboethics: The Applied Ethics for a New Science," in *Robot Ethics: The Ethical and Social Implications of Robotics*. Editors P. Lin, K. Abney, and G. A. Bekey (Cambridge, MA: MIT Press), 347–363.

Wales, J. (2020). "Empathy and Instrumentalization: Late Ancient Cultural Critique and the Challenge of Apparently Personal Robots," in *Culturally Sustainable Social Robotics: Proceedings of Robo-Philosophy 2020*. Editors J. Seibt, M. Nørskov, and O. S. Quick (Amsterdam: IOS Press), 114–124. doi:10.3233/faia200906

Walker, M. (2006). "Viewing Assignment of Moral Status to Service Robots from the Theological Ethics of Paul Tillich: Some Hard Questions," in AAAI Workshop Technical Report WS-06-09 (Menlo Park, California: The AAAI Press), 23–28. https://www.aaai.org/Library/Workshops/2006/ws06-09-005.php.

Wallach, W., and Allen., C. (2010). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Whitby, B. (2008). Sometimes It's Hard to Be a Robot: A Call for Action on the Ethics of Abusing Artificial Agents. *Interacting Comput.* 20 (3), 326–333. doi:10.1016/j.intcom.2008.02.002

Yancy, G., and Singer, P. (2015). Peter Singer: On Racism, Animal Rights and Human Rights. New York: *Opinionator*. October 8, 2015. Available at: https://opinionator.blogs.nytimes.com/2015/05/27/peter-singer-on-speciesism-and-racism/.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 08 November 2021
doi: 10.3389/frobt.2021.756242

The Conflict Between People's Urge to Punish AI and Legal Systems

Gabriel Lima[1,2], *Meeyoung Cha*[1,2]*, *Chihyung Jeon*[3] *and Kyung Sin Park*[4]

[1]*School of Computing, KAIST, Daejeon, South Korea,* [2]*Data Science Group, Institute for Basic Science, Daejeon, South Korea,* [3]*Graduate School of Science and Technology Policy, KAIST, Daejeon, South Korea,* [4]*School of Law, Korea University, Seoul, South Korea*

Regulating artificial intelligence (AI) has become necessary in light of its deployment in high-risk scenarios. This paper explores the proposal to extend legal personhood to AI and robots, which had not yet been examined through the lens of the general public. We present two studies (N = 3,559) to obtain people's views of electronic legal personhood vis-à-vis existing liability models. Our study reveals people's desire to punish automated agents even though these entities are not recognized any mental state. Furthermore, people did not believe automated agents' punishment would fulfill deterrence nor retribution and were unwilling to grant them legal punishment preconditions, namely physical independence and assets. Collectively, these findings suggest a conflict between the desire to punish automated agents and its perceived impracticability. We conclude by discussing how future design and legal decisions may influence how the public reacts to automated agents' wrongdoings.

Keywords: artificial intelligence, robots, AI, legal system, legal personhood, punishment, responsibility

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Henrik Skaug Sætra,
Østfold University College, Norway
Kamil Mamak,
Jagiellonian University, Poland

****Correspondence:***
Meeyoung Cha
mcha@ibs.re.kr

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

Received: *10 August 2021*
Accepted: *07 October 2021*
Published: *08 November 2021*

Citation:
Lima G, Cha M, Jeon C and Park KS (2021) The Conflict Between People's Urge to Punish AI and Legal Systems. Front. Robot. AI 8:756242. doi: 10.3389/frobt.2021.756242

1 INTRODUCTION

Artificial intelligence (AI) systems have become ubiquitous in society. To discover where and how these machines[1] affect people's lives does not require one to go very far. For instance, these automated agents can assist judges in bail decision-making and choose what information users are exposed to online. They can also help hospitals prioritize those in need of medical assistance and suggest who should be targeted by weapons during war. As these systems become widespread in a range of morally relevant environments, mitigating how their deployment could be harmful to those subjected to them has become more than a necessity. Scholars, corporations, public institutions, and nonprofit organizations have crafted several ethical guidelines to promote the responsible development of the machines affecting people's lives (Jobin et al., 2019). However, are ethical guidelines sufficient to ensure that such principles are followed? Ethics lacks the mechanisms to ensure compliance and can quickly become a tool for escaping regulation (Resseguier and Rodrigues, 2020). Ethics should not be a substitute for enforceable principles, and the path towards safe and responsible deployment of AI seems to cross paths with the law.

The latest attempt to regulate AI has been advanced by the European Union (EU; (European Commission, 2021)), which has focused on creating a series of requirements for high-risk systems (e.g., biometric identification, law enforcement). This set of rules is currently under public and

[1] We use the term "machine" as a interchangeable term for AI systems and robots, i.e., embodied forms of AI. Recent work on the human factors of AI systems has used this term to refer to both AI and robots (e.g., (Köbis et al., 2021)), and some of the literature that has inspired this research uses similar terms when discussing both entities, e.g., (Matthias, 2004).

scholarly scrutiny, and experts expect it to be the starting point of effective AI regulation. This research explores one proposal previously advanced by the EU that has received extensive attention from scholars but was yet to be studied through the lens of those most affected by AI systems, i.e., the general public. In this work, we investigate the possibility of extending legal personhood to autonomous AI and robots (Delvaux, 2017).

The proposal to hold machines, partly or entirely, liable for their actions has become controversial among scholars and policymakers. An open letter signed by AI and robotics experts denounced its prospect following the EU proposal (http://www.robotics-openletter.eu/). Scholars opposed to electronic legal personhood have argued that extending certain legal status to autonomous systems could create human liability shields by protecting humans from deserved liability (Bryson et al., 2017). Those who argue against legal personhood for AI systems regularly question how they could be punished (Asaro, 2011; Solaiman, 2017). Machines cannot suffer as punishment (Sparrow, 2007), nor do they have assets to compensate those harmed.

Scholars who defend electronic legal personhood argue that assigning liability to machines could contribute to the coherence of the legal system. Assigning responsibility to robots and AI could imbue these entities with realistic motivations to ensure they act accordingly (Turner, 2018). Some highlight that legal personhood has also been extended to other nonhumans, such as corporations, and doing so for autonomous systems may not be as implausible (Van Genderen, 2018). As these systems become more autonomous, capable, and socially relevant, embedding autonomous AI into legal practices becomes a necessity (Gordon, 2021; Jowitt, 2021).

We note that AI systems could be granted legal standing regardless of their ability to fulfill duties, e.g., by granting them certain rights for legal and moral protection (Gunkel, 2018; Gellers, 2020). Nevertheless, we highlight that the EU proposal to extend a specific legal status to machines was predicated on holding these systems legally responsible for their actions. Many of the arguments opposed to the proposal also rely on these systems' incompatibility with legal punishment and pose that these systems should not be granted legal personhood because they cannot be punished.

An important distinction in the proposal to extend legal personhood to AI systems and robots is its adoption under criminal and civil law. While civil law aims to make victims whole by compensating them (Prosser, 1941), criminal law punishes offenses. Rights and duties come in distinct bundles such that a legal person, for instance, may be required to pay for damages under civil law and yet not be held liable for a criminal offense (Kurki, 2019). The EU proposal to extend legal personhood to automated systems has focused on the former by defending that they could make "good any damage they may cause." However, scholarly discussion has not been restricted to the civil domain and has also inquired how criminal offenses caused by AI systems could be dealt with (Abbott, 2020).

Some of the possible benefits, drawbacks, and challenges of extending legal personhood to autonomous systems are unique to civil and criminal law. Granting legal personhood to AI systems may facilitate compensating those harmed under civil law (Turner, 2018), while providing general deterrence (Abbott, 2020) and psychological satisfaction to victims (e.g., through revenge (Mulligan, 2017)) if these systems are criminally punished. Extending civil liability to AI systems means these machines should hold assets to compensate those harmed (Bryson et al., 2017). In contrast, the difficulties of holding automated systems criminally liable extend to other domains, such as how to define an AI system's mind, how to reduce it to a single actor (Gless et al., 2016), and how to grant them physical independence.

The proposal to adopt electronic legal personhood addresses the difficult problem of attributing responsibility for AI systems' actions, i.e., the so-called responsibility gap (Matthias, 2004). Self-learning and autonomous systems challenge epistemic and control requirements for holding actors responsible, raising questions about who should be blamed, punished, or answer for harms caused by AI systems (de Sio and Mecacci, 2021). The deployment of complex algorithms leads to the "problem of many things," where different technologies, actors, and artifacts come together to complicate the search for a responsible entity (Coeckelbergh, 2020). These gaps could be partially bridged if the causally responsible machine is held liable for its actions.

Some scholars argue that the notion of a responsibility gap is overblown. For instance, Johnson (2015) has asserted that responsibility gaps will only arise if designers choose and argued that they should instead proactively take responsibility for their creations. Similarly, Sætra (2021) has argued that even if designers and users may not satisfy all requirements for responsibility attribution, the fact that they chose to deploy systems that they do not understand nor have control over makes them responsible. Other scholars view moral responsibility as a pluralistic and flexible process that can encompass emerging technologies (Tigard, 2020).

Danaher (2016) has made a case for a distinct gap posed by the conflict between the human desire for retribution and the absence of appropriate subjects of retributive punishment, i.e., the retribution gap. Humans look for a culpable wrongdoer deserving of punishment upon harm and justify their intuitions with retributive motives (Carlsmith and Darley, 2008). AI systems are not appropriate subjects of these retributive attitudes as they lack the necessary conditions for retributive punishment, e.g., culpability.

The retribution gap has been criticized by other scholars, who defend that people could exert control over their retributive intuitions (Kraaijeveld, 2020) and argue that conflicts between people's intuitions and moral and legal systems are dangerous only if they destabilize such institutions (Sætra, 2021). This research directly addresses whether such conflict is real and could pose challenges to AI systems' governance. Coupled with previous work finding that people blame AI and robots for harm (e.g., (Kim and Hinds, 2006; Malle et al., 2015; Furlough et al., 2021; Lee et al., 2021; Lima et al., 2021)), there seems to exist a clash between people's reactive attitudes towards harms caused by automated systems and their feasibility. This conflict is yet to be studied empirically.

We investigate this friction. We question whether people would punish AI systems in situations where human agents would typically be held liable. We also inquire whether these reactive attitudes can be grounded on crucial components of legal punishment, i.e., some of its requirements and functions.

Previous work on the proposal to extend legal standing to AI systems has been mostly restricted to the normative domain, and research is yet to investigate whether philosophical intuitions concerning the responsibility gap, retribution gap, and electronic legal personhood have similarities with the public view. We approach this research question as a form of experimental philosophy of technology (Kraaijeveld, 2021). This research does not defend that responsibility and retribution gaps are real or can be solved by other scholars' proposals. Instead, we investigate how people's reactive attitudes towards harms caused by automated systems may clash with legal and moral doctrines and whether they warrant attention.

Recent work has explored how public reactions to automated vehicles (AVs) could help shape future regulation (Awad et al., 2018). Scholars posit that psychology research could augment information available to policymakers interested in regulating autonomous machines (Awad et al., 2020a). This body of literature acknowledges that the public view should not be entirely embedded into legal and governance decisions due to harmful and irrational biases. Yet, they defend that obtaining the general public's attitude towards these topics can help regulators discern policy decisions and prepare for possible conflicts.

Viewing the issues of responsibility posed by automated systems as political questions, Sætra (2021) has defended that these questions should be subjected to political deliberation. Deciding how to attribute responsibility comes with inherent trade-offs that one should balance to achieve responsible and beneficial innovation. A crucial stakeholder in this endeavor is those who are subjected to the indirect consequences of widespread deployment of automated systems, i.e., the public (Dewey and Rogers, 2012). Scholars defend that automated systems "should be regulated according to the political will of a given community" (Sætra and Fosch-Villaronga, 2021), where the general public is a major player. Acknowledging the public opinion facilitates the political process to find common ground for the successful regulation of these new technologies. If legal responsibility becomes too detached from the folk conception of responsibility, the law might become unfamiliar to those whose behavior it aims to regulate, thus creating the "law in the books" instead of the "law in action" (Brożek and Janik, 2019).

People's expectations and preconceptions of AI systems and robots have several implications to their adoption, development, and regulation (Cave and Dihal, 2019). For instance, fear and hostility may hinder the adoption of beneficial technology (Cave et al., 2018; Bonnefon et al., 2020), whereas a more positive take on AI and robots may lead to unreasonable expectations and overtrust—which scholars have warned against (Bansal et al., 2019). Narratives about AI and robots also inform and open new directions for research among developers and shape the views of both policymakers and its constituents (Cave and Dihal, 2019). This research contributes to the maintenance of the "algorithmic social contract," which aims to embed societal values into the governance of new technologies (Rahwan, 2018). By understanding how all stakeholders involved in developing, deploying, and using AI systems react to these new technologies, those responsible for making governance decisions can be better informed of any existing conflicts.

2 METHODS

Our research inquired how people's moral judgments of automated systems may clash with existing legal doctrines through a survey-based study. We recruited 3,315 US residents through Amazon Mechanical Turk (see SI for demographic information), who attended a study where they 1) indicated their perception of automated agents' liability and 2) attributed responsibility, punishment, and awareness to a wide range of entities that could be held liable for harms caused by automated systems under existing legal doctrines.

We employed a between-subjects study design, in which each participant was randomly assigned to a scenario, an agent, and an autonomy level. Scenarios covered two environments where automated agents are currently deployed: medicine and war (see SI for study materials). Each scenario posited three agents: an AI program, a robot (i.e., an embodied form of AI), or a human actor. Although the proposal of extending legal standing to AI systems and robots have similarities, they also have distinct aspects worth noting. For instance, although a "robot death penalty" may be a viable option through its destruction, "killing" an AI system may not have the same expressive benefits due to varying levels of anthropomorphization. However, extensive literature discusses the two actors in parallel, e.g., (Turner, 2018; Abbott, 2020). We come back to this distinction in our final discussion. Finally, our study introduced each actor as either "supervised by a human" or "completely autonomous."

Participants assigned to an automated agent first evaluated whether punishing it would fulfill some of legal punishment's functions, namely reform, deterrence, and retribution (Solum, 1991; Asaro, 2007). They also indicated whether they would be willing to grant assets and physical independence to automated systems—two factors that are preconditions for civil and criminal liability, respectively. If automated systems do not hold assets to be taken away as compensation for those they harm, they cannot be held liable under civil law. Similarly, if an AI system or robot does not possess any level of physical independence, it becomes hard to imagine their criminal punishment. These questions were shown in random order and answered using a 5-point bipolar scale.

After answering this set of questions or immediately after consenting to the research terms for those assigned to a human agent, participants were shown the selected vignette in plain text. They were then asked to attribute responsibility, punishment, and awareness to their assigned agent. Responsibility and punishment are closely related to the proposal of adopting electronic legal personhood, while awareness plays a major role in legal judgments (e.g., mens rea in criminal law, negligence in civil law). We also identified a series of entities (hereafter associates) that could be held liable under existing legal doctrines, such as an automated system's manufacturer under product liability, and asked participants to attribute the same variables to each of them. All questions were answered using a 4-pt scale. Entities were shown in random order and one at a time.

We present the methodology details and study materials in the SI. A replication with a demographically representative sample ($N = 244$) is also shown in the SI to substantiate all of the findings presented in the main text. This research had been

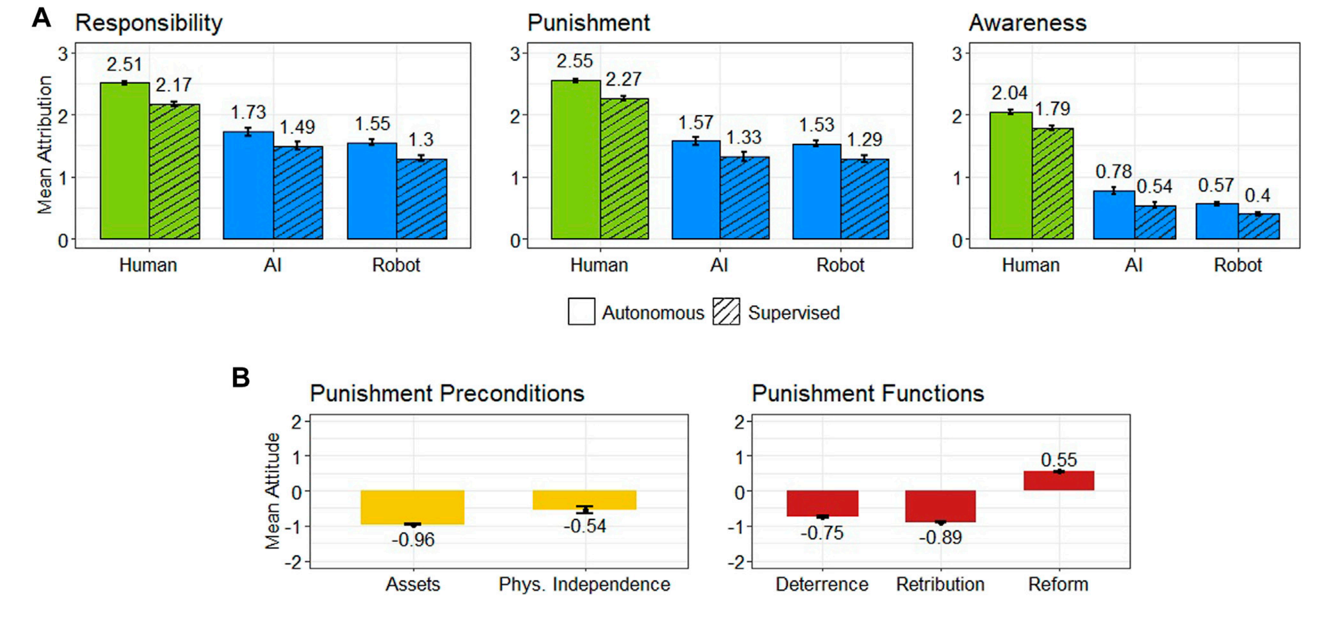

FIGURE 1 | Attribution of responsibility, punishment, and awareness to human agents, AI systems, and robots upon a legal offense **(A)**. Participants' attitudes towards granting legal punishment preconditions to AI systems and robots (e.g., assets and physical independence) and respondents' views that automated agents' punishment would (not) satisfy the deterrence, retributive, and reformative functions of legal punishment **(B)**. Standard errors are shown as error bars.

approved by the first author's Institutional Review Board (IRB). All data and scripts are available at the project's repository: https://bit.ly/3AMEJjB.

3 RESULTS

Figure 1A shows the mean values of responsibility and punishment attributed to each agent depending on their autonomy level. Automated agents were deemed moderately responsible for their harmful actions (M = 1.48, SD = 1.16), and participants wished to punish AI and robots to a significant level (M = 1.42, SD = 1.28). In comparison, human agents were held responsible (M = 2.34, SD = 0.83) and punished (M = 2.41, SD = 0.82) to a larger degree.

A 3 (agent: AI, robot, human) × 2 (autonomy: completely autonomous, supervised) ANOVA on participants' judgments of responsibility revealed main effects of both agent (F (2, 3309) = 906.28, $p < 0.001$, $\eta_p^2 = 0.35$) and autonomy level (F (1, 3309) = 43.84, $p < 0.001$, $\eta_p^2 = 0.01$). The extent to which participants wished to punish agents was also dependent on the agent (F (2, 3309) = 391.61, $p < 0.001$, $\eta_p^2 = 0.16$) and its autonomy (F (1, 3309) = 45.56, $p < 0.001$, $\eta_p^2 = 0.01$). The interaction between these two factors did not reach significance in any of the models ($p > 0.05$). Autonomous agents were overall viewed as more responsible and deserving of a larger punishment for their actions than their supervised counterparts. We did not observe noteworthy differences between AI systems and robots; the latter were deemed marginally less responsible than AI systems.

Figure 1A shows the mean perceived awareness of AI, robots, and human agents upon a legal offense. Participants perceived automated agents as only slightly aware of their actions (M = 0.54, SD = 0.88), while human agents were considered somewhat aware (M = 1.92, SD = 1.00). A 3 × 2 ANOVA model revealed main effects for both agent type (F (2, 3309) = 772.51, $p < 0.001$, η_p^2 = 0.35) and autonomy level (F (1, 3309) = 43.87, $p < 0.001$, $\eta_p^2 = 0.01$). The interaction between them was not significant (p = 0.401). Robots were deemed marginally less aware of their offenses than AI systems. A mediation analysis revealed that perceived awareness of AI systems (coded as -1) and robots (coded as 1) mediated judgments of responsibility (partial mediation, coef = −0.04, 95% CI [−0.06, −0.02]) and punishment (complete mediation, coef = −0.05, 95% CI [−0.07, −0.02]).

The leftmost plot of **Figure 1B** shows participants' attitudes towards granting assets and some level of physical independence to AI and robots using a 5-pt scale. These two concepts are crucial preconditions for imposing civil and criminal liability, respectively. Participants were largely contrary to allowing automated agents to hold assets (M = −0.96, SD = 1.16) or physical independence (M = −0.55, SD = 1.30). **Figure 1B** also shows the extent to which participants believed the punishment of AI and robots might satisfy deterrence, retribution, and reform, i.e., some of legal punishment's functions. Respondents did not believe punishing an automated agent would fulfill its retributive functions (M = −0.89, SD = 1.12) or deter them from future offenses (M = −0.75, SD = 1.22); however, AI and robots were viewed as able to learn from their wrongful actions (M = 0.55, SD = 1.17). We only observed marginal effects ($\eta_p^2 \leq$ 0.01) of agent type and autonomy in participants' attitudes towards preconditions and functions of legal punishment and present these results in the SI.

The viability and effectiveness of AI systems' and robots' punishment depend on fulfilling certain legal punishment's preconditions and functions. As discussed above, the incompatibility between legal punishment and automated

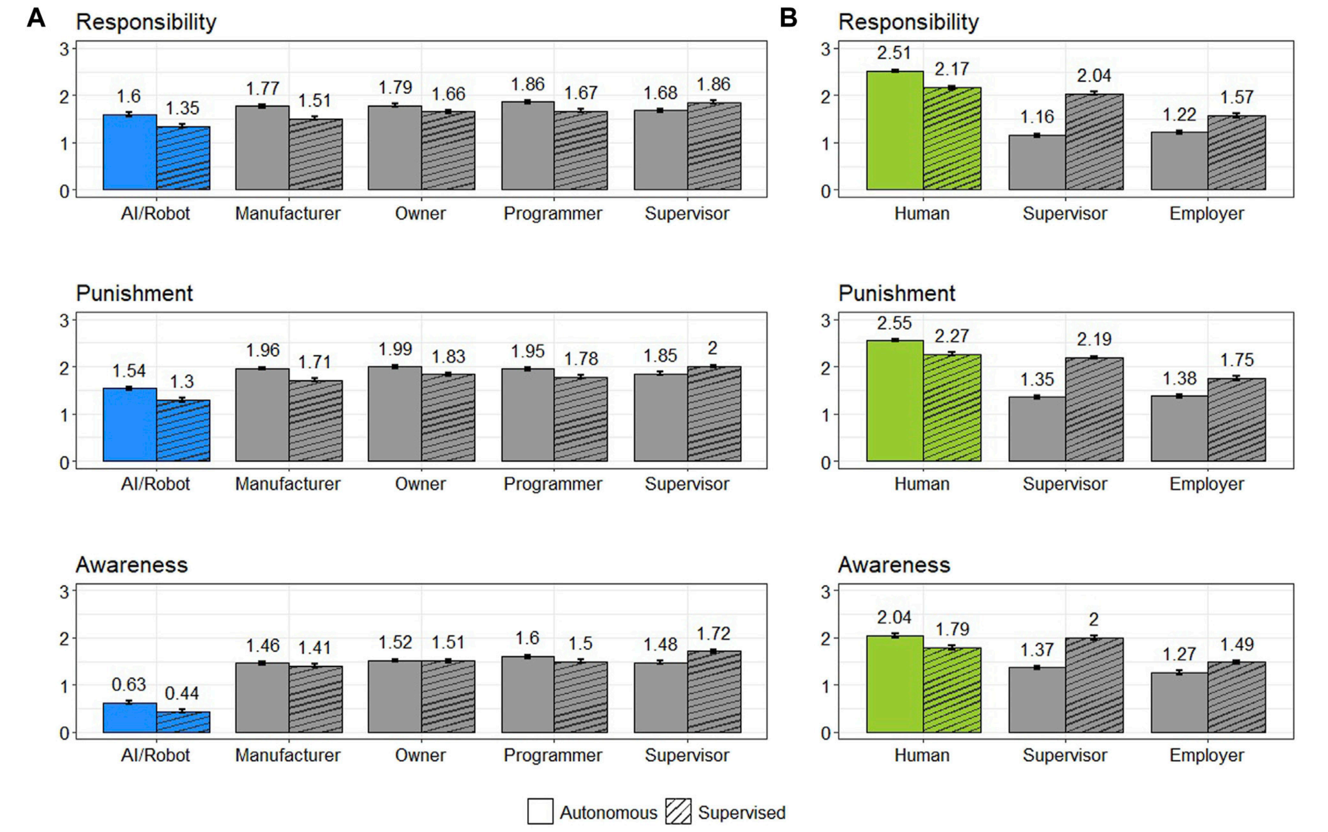

FIGURE 2 | Attribution of responsibility, punishment, and awareness to AI systems, robots, and entities that could be held liable under existing doctrines (i.e., associates) **(A)**. Assignment of responsibility, punishment, and awareness to human agents and corresponding associates **(B)**. Standard errors are shown as error bars.

agents is a common argument against the adoption of electronic legal personhood. Collectively, our results suggest a conflict between people's desire to punish AI and robots and the punishment's perceived effectiveness and feasibility.

We also observed that the extent to which participants wished to punish automated agents upon wrongdoing correlated with their attitudes towards granting them assets (r (1935) = 0.11, $p < 0.001$) and physical independence (r (224) = 0.21, $p < 0.001$). Those who anticipated the punishment of AI and robots to fulfill deterrence (r (1711) = 0.34, $p < 0.001$) and retribution (r (1711) = 0.28, $p < 0.001$) also tended to punish them more. However, participants' views concerning automated agents' reform were not correlated with their punishment judgments (r (1711) = –0.02, p = 0.44). In summary, more positive attitudes towards granting assets and physical independence to AI and robots were associated with larger punishment levels. Similarly, participants that perceived automated agents' punishment as more successful concerning deterrence and retribution punished them more. Nevertheless, most participants wished to punish automated agents regardless of the punishment's infeasibility and unfulfillment of retribution and deterrence.

Participants also judged a series of entities that could be held liable under existing liability models concerning their responsibility, punishment, and awareness for an agent's wrongful action. All of the automated agents' associates were judged responsible, deserving of punishment, and aware of the agents' actions to a similar degree (see **Figure 2**). The supervisor of a supervised AI or robot was judged more responsible, aware, and deserving of punishment than that of a completely autonomous system. In contrast, attributions of these three variables to all other associates were larger in the case of an autonomous agent. In the case of human agents, their employers and supervisors were deemed more responsible, aware, and deserving of punishment when the actor was supervised. We present a complete statistical analysis of these results in the SI.

4 DISCUSSION

Our findings demonstrate a conflict between participants' desire to punish automated agents for legal offenses and their perception that such punishment would not be successful in achieving deterrence or retribution. This clash is aggravated by participants' unwillingness to grant AI and robots what is needed to legally punish them, i.e., assets for civil liability and physical independence for criminal liability. This contradiction in people's moral judgments suggests that people wish to punish AI and robots even though they believe that doing so would not be successful, nor are they willing to make it legally viable.

These results are in agreement with Danaher's (2016) retribution gap. Danaher acknowledges that people might blame and punish AI and robots for wrongful behavior due to humans' retributive nature, although they may be wrong in doing so. Our data implies that Danaher's concerns about the retribution gap are significant and can be extended to other considerations, i.e., deterrence and the preconditions for legal punishment. Past research shows that people also ground their punishment judgments in functions other than retribution (Twardawski et al., 2020). Public intuitions concerning the punishment of automated agents are even more contradictory than previously advanced by Danaher: they wish to punish AI and robots for harms even though their punishment would not be successful in achieving some of legal punishment's functions or even viable, given that people would not be willing to grant them what is necessary to punish them.

Our results show that even if responsibility and retribution gaps can be easily bridged as suggested by some scholars (Sætra, 2021; Tigard, 2020; Johnson, 2015), there still exists a conflict between the public reaction to harms caused by automated systems and their moral and legal feasibility. The public is an important stakeholder in the political deliberation necessary for the beneficial regulation of AI and robots, and their perspective should not be rejected without consideration. An empirical question that our results pose is whether this conflict warrants attention from scholars and policymakers, i.e., if they destabilize political and legal institutions (Sætra, 2021) or leads to lack of trust in legal systems (Abbott, 2020). For instance, it may well be that the public may need to be taught to exert control over their moral intuitions, as suggested by Kraaijeveld (2020).

Although participants did not believe punishing an automated agent would satisfy the retributive and deterrence aspects of punishment, they viewed robots and AI systems as capable of learning from their mistakes. Reform may be the crucial component of people's desire to punish automated agents. Although the current research might not be able to clear this inquiry, we highlight that future work should explore how participants imagine the reform of automated agents. Reprogramming an AI system or robots can prevent future offenses, yet it will not satisfy other indirect reformative functions of punishment, e.g., teaching others that a specific action is wrong. Legal punishment, as it stands, does not achieve the reprogramming necessary for AI and robots. Future studies may question how people's preconceptions of automated agents' reprogramming influence people's moral judgments.

It might be argued that our results are caused by how the study was constructed. For instance, participants who punished automated agents might have reported being more optimistic about its feasibility so that their responses become compatible. However, we observe trends that methodological biases cannot explain but can only result from participants' a priori contradiction (see SI for detailed methodology). This work does not posit this contradiction as a universal phenomenon; we observed a significant number of participants attributing no punishment whatsoever to electronic agents. Nonetheless, we observed similar results in a demographically representative sample of respondents (see SI).

We did not observe significant differences between punishment judgments of AI systems and robots. The differences in responsibility and awareness judgments were marginal and likely affected by our large sample size. As discussed above, there are different challenges when adopting electronic legal personhood for AI and robots. Embodied machines may be easier to punish criminally if legal systems choose to do so, for instance through the adoption of a "robot death penalty." Nevertheless, our results suggest that the conflict between people's moral intuitions and legal systems may be independent of agent type. Our study design did not control for how people imagined automated systems, which could have affected how people make moral judgments about machines. For instance, previous work has found that people evaluate the moral choices of a human-looking robot as less moral than humans' and non-human robots' decisions (Laakasuo et al., 2021).

People largely viewed AI and robots as unaware of their actions. Much human-computer interaction research has focused on developing social robots that can elicit mind perception through anthropomorphization (Waytz et al., 2014; Darling, 2016). Therefore, we may have obtained higher perceived awareness had we introduced what the robot or AI looked like, which in turn could have affected respondents' responsibility and punishment judgments, as suggested by Bigman et al. (2019) and our mediation analysis. These results may also vary by actor, as robots are subject to higher levels of anthropomorphization. Past research has also shown that if an AI system is described as an anthropomorphized agent rather than a mere tool, it is attributed more responsibility for creating a painting (Epstein et al., 2020). A similar trend was observed with autonomous AI and robots, which were assigned more responsibility and punishment than supervised agents, as previously found in the case of autonomous vehicles (Awad et al., 2020b) and other scenarios (Kim and Hinds, 2006; Furlough et al., 2021).

4.1 The Importance of Design, Social, and Legal Decisions

Participants' attitudes concerning the fulfillment of punishment preconditions and functions by automated agents were correlated with the extent to which respondents wished to punish AI and robots. This finding suggests that people's moral judgments of automated agents' actions can be nudged based on how their feasibility is introduced.

For instance, to clarify that punishing AI and robots will not satisfy the human need for retribution, will not deter future offenses, or is unviable given they cannot be punished similarly to other legal persons may lead people to denounce automated agents' punishment. If legal and social institutions choose to embrace these systems, e.g., by granting them certain legal status, nudges towards granting them certain perceived independence or private property may affect people's decision to punish them. Future work should delve deeper into the causal relationship between people's attitudes towards the topic and their attribution of punishment to automated agents.

Our results highlight the importance of design, social, and legal decisions in how the general public may react to automated agents. Designers should be aware that developing systems that are perceived as aware by those interacting with them may lead to heightened moral judgments. For instance, the benefits of automated agents may be nullified if their adoption is impaired by unfulfilled perceptions that these systems should be punished. Legal decisions concerning the regulation of AI and their legal standing may also influence how people react to harms caused by automated agents. Social decisions concerning how to insert AI and robots into society, e.g., as legal persons, should also affect how we judge their actions. Future decisions should be made carefully to ensure that laypeople's reactions to harms caused by automated systems do not clash with regulatory efforts.

5 CONCLUDING REMARKS

Electronic legal personhood grounded on automated agents' abilities to fulfill duties does not seem a viable path towards the regulation of AI. This approach can only become an option if AI and robots are granted assets or physical independence, which would allow civil or criminal liability to be imposed, or if punishment functions and methods are adapted to AI and robots. People's intuitions about automated agents' punishment are somewhat similar to scholars who oppose the proposal. However, a significant number of people still wish to punish AI and robots independently of their a priori intuitions.

By no means this research proposes that robots and AI should be the sole entities to hold liability for their actions. In contrast, responsibility, awareness, and punishment were assigned to all associates. We thus posit that distributing liability among all entities involved in deploying these systems would follow the public perception of the issue. Such a model could take joint and several liability models as a starting point by enforcing the proposal that various entities should be held jointly liable for damages.

Our work also raises the question of whether people wish to punish AI and robots for reasons other than retribution, deterrence, and reform. For instance, the public may punish electronic agents for general or indirect deterrence (Twardawski et al., 2020). Punishing an AI could educate humans that a specific action is wrong without the negative consequences of human punishment. Recent literature in moral psychology also proposes that humans might strive for a morally coherent world, where seemingly contradictory judgments arise so that the public perception of agents' moral qualities match the moral qualities of their actions' outcomes (Clark et al., 2015). We highlight that legal punishment is not only directed at the wrongdoer but also fulfills other functions in society that future work should inquire about when dealing with automated agents. Finally, our work poses the question of whether proactive actions towards holding existing legal persons liable for harms caused by automated agents would compensate for people's desire to punish them. For instance, future work might examine whether punishing a system's manufacturer may decrease the extent to which people punish AI and robots. Even if the responsibility gap can be easily solved, conflicts between the public and legal institutions might continue to pose challenges to the successful governance of these new technologies.

We selected scenarios from active areas of AI and robotics (i.e., medicine and war; see SI). People's moral judgments might change depending on the scenario or background. The proposed scenarios did not introduce, for the sake of feasibility and brevity, much of the background usually considered when judging someone's actions legally. We did not control for any previous attitudes towards AI and robots or knowledge of related areas, such as law and computer science, which could result in different judgments among the participants.

This research has found a contradiction in people's moral judgments of AI and robots: they wish to punish automated agents, although they know that doing so is not legally viable nor successful. We do not defend the thesis that automated agents should be punished for legal offenses or have their legal standing recognized. Instead, we highlight that the public's preconceptions of AI and robots influence how people react to their harmful consequences. Most crucially, we showed that people's reactions to these systems' failures might conflict with existing legal and moral systems. Our research showcases the importance of understanding the public opinion concerning the regulation of AI and robots. Those making regulatory decisions should be aware of how the general public may be influenced or clash with such commitments.

DATA AVAILABILITY STATEMENT

The datasets and scripts used for analysis in this study can be found at https://bitly.com/3AMEJjB.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board (IRB) at KAIST. The patients/participants provided their informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors designed the research. GL conducted the research. GL analyzed the data. GL wrote the paper, with edits from MC, CJ, and KS.

FUNDING

This research was supported by the Institute for Basic Science (IBS-R029-C2).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.756242/full#supplementary-material

REFERENCES

Abbott, R. (2020). *The Reasonable Robot: Artificial Intelligence and the Law*. Cambridge University Press.

Asaro, P. M. (2011). 11 a Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. *Robot Ethics ethical Soc. implications robotics*, 169–186.

Asaro, P. M. (2007). Robots and Responsibility from a Legal Perspective. *Proc. IEEE* 4, 20–24.

Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2020a). Crowdsourcing Moral Machines. *Commun. ACM* 63, 48–55. doi:10.1145/3339904

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The Moral Machine experiment. *Nature* 563, 59–64. doi:10.1038/s41586-018-0637-6

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., et al. (2020b). Drivers Are Blamed More Than Their Automated Cars when Both Make Mistakes. *Nat. Hum. Behav.* 4, 134–143. doi:10.1038/s41562-019-0762-8

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019). "Beyond Accuracy: The Role of Mental Models in Human-Ai Team Performance," in Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2–11.

Bigman, Y. E., Waytz, A., Alterovitz, R., and Gray, K. (2019). Holding Robots Responsible: The Elements of Machine Morality. *Trends Cognitive Sciences* 23, 365–368. doi:10.1016/j.tics.2019.02.008

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2020). *The Moral Psychology of AI and the Ethical Opt-Out Problem*. Oxford, UK: Oxford University Press.

Brożek, B., and Janik, B. (2019). Can Artificial Intelligences Be Moral Agents. *New Ideas Psychol.* 54, 101–106. doi:10.1016/j.newideapsych.2018.12.002

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: the Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25, 273–291. doi:10.1007/s10506-017-9214-9

Carlsmith, K. M., and Darley, J. M. (2008). Psychological Aspects of Retributive justice. *Adv. Exp. Soc. Psychol.* 40, 193–236. doi:10.1016/s0065-2601(07)00004-4

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., et al. (2018). *Portrayals and Perceptions of Ai and Why They Matter*.

Cave, S., and Dihal, K. (2019). Hopes and Fears for Intelligent Machines in Fiction and Reality. *Nat. Mach Intell.* 1, 74–78. doi:10.1038/s42256-019-0020-9

Clark, C. J., Chen, E. E., and Ditto, P. H. (2015). Moral Coherence Processes: Constructing Culpability and Consequences. *Curr. Opin. Psychol.* 6, 123–128. doi:10.1016/j.copsyc.2015.07.016

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Sci. Eng. Ethics* 26, 2051–2068. doi:10.1007/s11948-019-00146-8

Danaher, J. (2016). Robots, Law and the Retribution gap. *Ethics Inf. Technol.* 18, 299–309. doi:10.1007/s10676-016-9403-3

Darling, K. (2016). "Extending Legal protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects," in *Robot Law* (Edward Elgar Publishing).

de Sio, F. S., and Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them. *Philos. Tech.*, 1–28. doi:10.1007/s13347-021-00450-x

Delvaux, M. (2017). Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (Inl)). European Parliament Committee on Legal Affairs.

Dewey, J., and Rogers, M. L. (2012). *The Public and Its Problems: An Essay in Political Inquiry*. Penn State Press.

Epstein, Z., Levine, S., Rand, D. G., and Rahwan, I. (2020). Who Gets Credit for Ai-Generated Art. *Iscience* 23, 101515. doi:10.1016/j.isci.2020.101515

European Commission (2021). Proposal for a Regulation of the European Parliament and of the council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain union Legislative Acts).

Furlough, C., Stokes, T., and Gillan, D. J. (2021). Attributing Blame to Robots: I. The Influence of Robot Autonomy. *Hum. Factors* 63, 592–602. doi:10.1177/0018720819880641

Gellers, J. C. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law (Edition 1)*. Routledge.

Gless, S., Silverman, E., and Weigend, T. (2016). If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability. *New Criminal L. Rev.* 19, 412–436. doi:10.1525/nclr.2016.19.3.412

Gordon, J. S. (2021). Artificial Moral and Legal Personhood. *AI Soc.* 36, 457–471. doi:10.1007/s00146-020-01063-2

Gunkel, D. J. (2018). *Robot Rights*. mit Press.

Jobin, A., Ienca, M., and Vayena, E. (2019). The Global Landscape of Ai Ethics Guidelines. *Nat. Mach Intell.* 1, 389–399. doi:10.1038/s42256-019-0088-2

Johnson, D. G. (2015). Technology with No Human Responsibility. *J. Bus Ethics* 127, 707–715. doi:10.1007/s10551-014-2180-1

Jowitt, J. (2021). Assessing Contemporary Legislative Proposals for Their Compatibility With a Natural Law Case for AI Legal Personhood. *AI Soc.* 36, 499–508. doi:10.1007/s00146-020-00979-z

Kim, T., and Hinds, P. (2006). "Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction," in ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication (IEEE), 80–85.

Köbis, N., Bonnefon, J.-F., and Rahwan, I. (2021). Bad Machines Corrupt Good Morals. *Nat. Hum. Behav.* 5, 679–685. doi:10.1038/s41562-021-01128-2

Kraaijeveld, S. R. (2020). Debunking (The) Retribution (gap). *Sci. Eng. Ethics* 26, 1315–1328. doi:10.1007/s11948-019-00148-6

Kraaijeveld, S. R. (2021). Experimental Philosophy of Technology. *Philos. Tech.*, 1–20. doi:10.1007/s13347-021-00447-6

Kurki, V. A. (2019). *A Theory of Legal Personhood*. Oxford University Press.

Laakasuo, M., Palomäki, J., and Köbis, N. (2021). Moral Uncanny valley: a Robot's Appearance Moderates How its Decisions Are Judged. *Int. J. Soc. Robotics*, 1–10. doi:10.1007/s12369-020-00738-6

Lee, M., Ruijten, P., Frank, L., de Kort, Y., and IJsselsteijn, W. (2021). "People May Punish, but Not Blame Robots," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–11. doi:10.1145/3411764.3445284

Lima, G., Grgić-Hlača, N., and Cha, M. (2021). "Human Perceptions on Moral Responsibility of Ai: A Case Study in Ai-Assisted Bail Decision-Making," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–17. doi:10.1145/3411764.3445260

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). "Sacrifice One for the Good of many? People Apply Different Moral Norms to Human and Robot Agents," in 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), 117–124.

Matthias, A. (2004). The Responsibility gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics Inf. Technol.* 6, 175–183. doi:10.1007/s10676-004-3422-1

Mulligan, C. (2017). Revenge against Robots. *SCL Rev.* 69, 579.

Prosser, W. L. (1941). *Handbook of the Law of Torts*. West Publishing.

Rahwan, I. (2018). Society-in-the-loop: Programming the Algorithmic Social Contract. *Ethics Inf. Technol.* 20, 5–14. doi:10.1007/s10676-017-9430-8

Resseguier, A., and Rodrigues, R. (2020). Ai Ethics Should Not Remain Toothless! a Call to Bring Back the Teeth of Ethics. *Big Data Soc.* 7, 2053951720942541. doi:10.1177/2053951720942541

Sætra, H. S. (2021). Confounding Complexity of Machine Action: a Hobbesian Account of Machine Responsibility. *Int. J. Technoethics (Ijt)* 12, 87–100. doi:10.4018/IJT.20210101.oa1

Sætra, H. S., and Fosch-Villaronga, E. (2021). Research in Ai Has Implications for Society: How Do We Respond. *Morals & Machines* 1, 60–73. doi:10.5771/2747-5174-2021-1-60

Solaiman, S. M. (2017). Legal Personality of Robots, Corporations, Idols and Chimpanzees: a Quest for Legitimacy. *Artif. Intell. L.* 25, 155–179. doi:10.1007/s10506-016-9192-3

Solum, L. B. (1991). Legal Personhood for Artificial Intelligences. *NCL Rev.* 70, 1231.

Sparrow, R. (2007). Killer Robots. *J. Appl. Philos.* 24, 62–77. doi:10.1111/j.1468-5930.2007.00346.x

Tigard, D. W. (2020). There Is No Techno-Responsibility gap. *Philos. Tech.*, 1–19. doi:10.1007/s13347-020-00414-7

Turner, J. (2018). *Robot Rules: Regulating Artificial Intelligence*. Springer.

Twardawski, M., Tang, K. T. Y., and Hilbig, B. E. (2020). Is it All about Retribution? the Flexibility of Punishment Goals. *Soc. Just Res.* 33, 195–218. doi:10.1007/s11211-020-00352-x

van den Hoven van Genderen, R. (2018). “Do we Need New Legal Personhood in the Age of Robots and Ai,” in *Robotics, AI and the Future of Law* (Springer), 15–55. doi:10.1007/978-981-13-2874-9_2

Waytz, A., Heafner, J., and Epley, N. (2014). The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle. *J. Exp. Soc. Psychol.* 52, 113–117. doi:10.1016/j.jesp.2014.01.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 22 November 2021
doi: 10.3389/frobt.2021.768092

Robot Responsibility and Moral Community

Dane Leigh Gogoshin *

Department of Practical Philosophy, RADAR Research Group, Faculty of Social Sciences, University of Helsinki, Helsinki, Finland

It is almost a foregone conclusion that robots cannot be morally responsible agents, both because they lack traditional features of moral agency like consciousness, intentionality, or empathy and because of the apparent senselessness of holding them accountable. Moreover, although some theorists include them in the moral community as moral patients, on the Strawsonian picture of moral community as requiring moral responsibility, robots are typically excluded from membership. By looking closely at our actual moral responsibility practices, however, I determine that the agency reflected and cultivated by them is limited to the kind of moral agency of which some robots are capable, not the philosophically demanding sort behind the traditional view. Hence, moral rule-abiding robots (if feasible) can be sufficiently morally responsible and thus moral community members, despite certain deficits. Alternative accountability structures could address these deficits, which I argue ought to be in place for those existing moral community members who share these deficits.

Keywords: moral responsibility, artificial moral agency, human-robot interaction, artificial intelligence, accountability structures

Edited by:
David Gunkel,
Northern Illinois University, United States

Reviewed by:
Sven Nyholm,
Utrecht University, Netherlands
Kamil Mamak,
Jagiellonian University, Poland

****Correspondence:***
Dane Leigh Gogoshin
dane.gogoshin@helsinki.fi

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

Received: *31 August 2021*
Accepted: *08 October 2021*
Published: *22 November 2021*

Citation:
Gogoshin DL (2021) Robot Responsibility and Moral Community. Front. Robot. AI 8:768092. doi: 10.3389/frobt.2021.768092

1 INTRODUCTION

Since P. F. Strawson's landmark essay, "Freedom and Resentment" (Strawson, 2008), morally responsible agency is taken to be a matter of being a fitting target of our responsibility practices.[1] What exactly this fittingness consists in varies by account, but in most basic terms, per Strawson (see also Wallace, 1994), it is an agent's capacity to fulfill society's basic normative demands and expectations. This capacity is instantiated in the practices of being held to account when we transgress or exceed, respectively, these demands and expectations. Our actual practices are thus taken as reflections of this capacity – i.e., of responsible agency. On my analysis (see also Gogoshin, 2020), these standards are much lower than those we traditionally associate with human moral agency or the standards which human agents are, in principle, capable of meeting. Rather than requiring robust moral reasons-responsiveness or autonomy, these practices require only sensitivity to them (a sensitivity to the sting of moral disapproval, condemnation, blame and punishment and to the pleasure of moral approval, praise,

[1] By moral responsibility practices, I mean moral approbation and disapprobation, praise, blame, sanction and reward and the reactive attitudes (e.g., resentment, indignation, love, gratitude, etc.).

and reward). In turn, they reflect and cultivate a limited kind of moral agency, one concerned with performance – behavior that conforms to moral values – not with "what's going on on the inside" (agents' reasons and intentions).

Should one have the capacity to reliably behave in accordance with the normative demands and expectations of one's social environment, one is thus morally responsible. On this basis, I argue that autonomous robots[2] (henceforth just "robots") who have the capacity to reliably behave in accordance with the relevant moral rules and values of their social environment (henceforth "moral rule-abiding robots")[3] are morally responsible agents. As a consequence, on the view that moral community membership is a matter of morally responsible agency (Strawson, 2008; Darwall, 2006)[4], such robots are moral community members too.[5] If this result is objectionable, then we ought to add further conditions to moral community membership than morally responsible agency (e.g., sentience[6]). If, however, we retain responsibility as a necessary condition of moral community, by defining it in any more demanding terms than those I lay out in this paper, we would likely have to reject many current members from our moral communities.[7]

This conception of moral agency is clearly in tension with a deeper, more substantive conception of morally responsible agency[8] – the one at stake in the free will debate – which is rooted in concerns about fairness and desert, for our identity as responsible, rational and in some way free agents (Holroyd, 2007), and for our ultimate moral aspirations. After all, only one who meets certain epistemic and control conditions and/or can meaningfully identify with their actions or attitudes can be blame- or praiseworthy. Furthermore, we value the capacity to recognize and respond to moral reasons. However, this level of agency is neither reflected nor cultivated by our responsibility practices. Accordingly, morally responsible agency falls short of full-blown, autonomous moral agency. I hypothesize that it is obtained, when it is, through a multiplicity of other factors which lie outside of the moral responsibility system. However, in order to meet the basic demands of morality and to function as a moral community (at least in the way that we do), this level of moral agency appears to be unnecessary and, what's more, given that the other factors behind our moral development are likely non-ubiquitous and contingent (i.e., dependent on one's environment, upbringing, socioeconomic status, cultural influences, education level, etc.), too demanding.

I proceed as follows. In **Section 2**, I situate my approach within the existing artificial moral agency debate. In **Sections 3** and **4**, I present my analysis of the moral responsibility practices, showing that 1) rather than agents' reasons for action, they reflect agents' capacity to comply with moral norms, and 2) insofar as they are regulative, they are largely conditioning practices which are limited to regulating behavior. I identify the limitations on moral agency of behavioral regulation. In **Section 5**, I argue that moral rule-abiding robots can meet the behavioral level of moral agency required for moral community membership and offer some additional reasons to support their membership. In **Section 6**, I explore potential objections to this argument and offer some solutions. I conclude in **Section 7**.

2 SITUATING THE PROPOSED VIEW

Although many theorists hold onto the traditional conception of full-blown moral agency as being a matter of moral responsibility

[2] Though there may be other relevant artificially intelligent systems, I limit my argument to robots that meet Sullins' definition of autonomous robots (Sullins, 2011: 154). "Autonomous" here refers to the roboticist or engineering sense in which Sullins uses it (see also Arkin, 2009).

[3] A matter which, admittedly, remains far from settled; see Sharkey (2020) for a sobering overview of the current debate. Ron Arkin (2009) makes the most confident case for robot ethicality; see also Nadeau (2006). I address this further toward the end of **Section 5**.

[4] According to Strawson (2008: 17), moral responsibility is a precondition of being "a term of moral relationships" and a moral community member. Zimmerman (2016: 251) states that Strawson takes all three concepts as synonymous. Per Darwall (2006: 17), moral responsibility is being subject to the moral reactive attitudes, which "presuppose the authority to demand and hold one another responsible for compliance with moral obligations (which just are the standards to which we can warrantedly hold each other as members of the moral community)."

[5] Among current technologies, self-driving vehicles come quite close to the robots I have in mind. They operate rather reliably in high stakes settings. I envision care robots as an imminent example. However, it is not clear whether there are any extant robots that meet all the relevant moral demands of morally impactful social roles or, for that matter, how wide the social context they are capable of performing reliably in should be in order to qualify as moral community members. High stakes social institutions require specialized skills and security clearance, from financial institutions, to legal (courtrooms, prisons, government offices), military, medical, educational, safety (e.g., land/air/sea traffic control), etc. institutions. As a result, most humans have highly limited access to society, but this does not preclude their moral community membership. Hence, what is to be understood by moral community is the community of responsible agents. The fact that moral rule-abiding robots could qualify (under my proposed view) as responsible agents provides an impetus to investigate the concept of moral community carefully and to make prescriptive claims which uphold our ultimate moral values. The present proposal does not perform this task, though it will present (in **Section 5**) some normative reasons why its conception of responsible agency might be a sufficient condition for community membership.

[6] On the proposed view, responsible agents need not be moral patients. However, we might wish to make moral patiency a requirement for moral community membership.

[7] See Gogoshin (2020) for a condensed version of the stronger argument—that moral rule-abiding robots are ideal moral agents per the moral responsibility system.

[8] I subsequently refer to this conception as "robust moral responsibility" or "substantive responsibility." I hold that it requires, *inter alia*, robust moral reasons-responsiveness, i.e., the ability to recognize and respond to moral considerations in a wide range of circumstances. To be substantively or robustly morally responsible is to be largely morally autonomous: governed/motivated by the moral reason directly. Compare also the Aristotelian ideal of the virtuous person.

(e.g., Sparrow, 2007; Parthemore and Whitby, 2013; Hakli and Mäkelä, 2019), thereby denying robots full-blown moral agency, in the words of Wendell Wallach (Wallach and Allen, 2009), artificial moral agents are necessary and inevitable. Since Floridi and Sanders (2004), there has been a growing trend to divorce the question of moral agency from moral responsibility specifically and from philosophical personhood more generally (see also Sullins, 2006), in order to expand the set of moral agents. This move eliminates distinctly human capacities such as consciousness from the necessary conditions of moral agency.[9] It is thus generally thought that robots cannot be responsible in the way that mature, neurotypical humans are. Along with recent proposals by Christian List (2021) and Daniel Tigard (2021), which I will address at the end of this section, the present proposal challenges that notion.

The current state of the artificial moral agency debate is laid out in detail in Behdadi and Munthe (2020); I will not attempt to reconstruct it here. As they note, the debate is largely divided into two approaches – the standard or traditional (cf. Johnson, 2006) and the functionalist (cf. Floridi and Sanders, 2004).[10] The first seeks to identify features of traditional moral agency and to determine whether robots might have them. The second seeks to identify whether the functions of moral agency can be fulfilled by robots. According to Behdadi and Munthe, these two views are rife with conceptual confusion and are hopelessly irreconcilable. They propose shifting the debate away from a determination of whether machines are moral agents and toward which, whether and to what extent they should become part of society.[11]

There are two alternative approaches of particular relevance to my proposed account – those of Mark Coeckelbergh (2009) and John Danaher (2020). They look to see whether robots could be the fitting targets – in some way – of our existing social practices as they relate to moral patiency, agency, or responsibility. They then take facts about those practices, namely that they are necessarily blind to agents' mental states (see Himma, 2009 re. the "other minds problem"), and conclude that robots who elicit these practices (responses) are fitting targets of them. This insight does not allow us to say that robots are thereby moral agents or morally responsible, or that they are fitting targets of the full range of our practices, or that they can fulfill all the functions which we tend to ascribe to mature, neurotypical human beings, however. Unlike my proposed account, it does not reveal the kind of moral competence to which our practices are sensitive.

Danaher (2020) prescribes ethical behaviorism, according to which we ought to attribute moral status to a robot if it behaves in a way that we interpret as a feature of those to whom we already ascribe moral status. If the capacity for suffering is grounds for moral status and a robot appears to be suffering, then we ought to attribute moral status to the robot. Since we attribute moral status to human beings on the basis of mental states which we can only infer from their behavioral representations, we ought to do so, Danaher argues, with humanoid robots. I disagree with Danaher's normative stance; however, ethical behaviorism is an approach which respects our epistemic limits. Even if there are mental states that provide the ultimate metaphysical grounds for our ethical principles, we can only know them by way of their behavioral representations (Danaher, 2020: 2028). This is reflected in our social practices and especially in our tendency to anthropomorphize other beings and entities.

These practices – even when they err on the side of caution toward the agent in question (better to treat someone/something well just in case it is sentient or conscious, etc.) – come with risks. For one, we risk expending our resources on those who cannot reciprocate them and for another, we are vulnerable to malicious deception, e.g., something that emulates pain could lure in and harm an unsuspecting good doer. There are other normative reasons (as pointed out in Darling, 2016 and Coeckelbergh, 2021) to avoid destructive behavior toward robots – human agent-centered reasons (relating to how our behavior affects our own character or moral worth) – but Danaher's approach captures something descriptively significant about our practices; we judge others based on very limited and fallible inferences. The reason that supports doing so with non-humans is that it appears to be our only means of ascertaining the morally relevant information.

Mark Coeckelbergh's earlier proposal (Coeckelbergh, 2009) of virtual agency and responsibility falls along similar lines. Coeckelbergh takes our existing social practices of ascribing these concepts to others as his theoretical starting point, observing that within human interactions, we ascribe agency and responsibility independently of "the real" (Coeckelbergh, 2009: 184). They are in this sense virtual concepts; we ascribe them to others on the basis of how we experience them and how they appear to us. Since we engage in these virtual ascriptions with humans and animals – often going so far as to attribute a will to the latter – we will (and do) and further, should engage likewise with robots. Since moral responsibility and agency are, as far as our means of assessing them goes, matters of appearance and performance, Coeckelbergh argues that our ascriptions of these concepts or features to robots ought also to be a matter of appearance and performance.

My proposed account follows what I take to be a related yet distinct approach. Rather than starting with a particular conception of moral agency as per the standard view, or investigating solely whether robots could fulfill the functions we ascribe to morally responsible agents, or as do Danaher and Coeckelbergh – taking our practices themselves as the basis for a prescriptive account of artificial moral agency – I utilize the Strawsonian methodology by taking our responsibility practices as a starting point and identifying the criteria at work in them, in order to determine the features of a morally responsible agent. Like Danaher and Coeckelbergh, I note that our practices are limited to behavioral assessments and as such, they do not sufficiently capture or reflect all morally relevant mental content.

However, I do not take our practices to settle the matter about moral agency which, like Peter Asaro (2006), I consider to be a

[9] See Champagne and Tonkens, 2015 and Himma, 2009; they argue that consciousness is only needed for moral responsibility (Behdadi and Munthe, 2020).

[10] With some exceptions; see Behdadi and Munthe (2020: 199–200).

[11] This seems right headed, for even if a clear determination is made that foreseeable robots cannot meet the criteria for moral agency that we take human beings to meet, or that robots cannot fulfill certain normative expectations, we are still faced with this question.

scalar phenomenon – with moral autonomy on the upper end of the spectrum. I take them to set the standards for morally responsible agency, which I take to be lower than for moral autonomy. I hold that we ought to adopt further practices (and revise existing ones where possible[12]), in order to cultivate moral autonomy – something I do not see as a realistic (or desirable[13]) goal for robot design. Although certain robots are in principle capable of passing the performance-based test for morally responsible agency, I do not argue that this is a sufficient basis for the rights and privileges that other moral community members may be owed in virtue of other features such as sentience or personhood. So although this approach provides an answer to whether robots pass the Strawsonian requirements for moral community membership, it does not investigate whether this is in fact a desirable outcome. I will, however, put forward a conception of morality (in **Section 5**) as well as practical reasons (throughout), which offer practice-independent support for behavioral moral agency as morally responsible agency and as a potentially sufficient basis for moral community membership.

Before moving on, however, it is important to situate my proposal with respect to the other recent accounts of artificial moral responsibility previously mentioned (List, 2021; Tigard, 2021). Christian List argues that certain artificial intelligent systems, like certain group agents (e.g., corporations), who meet the following conditions on responsible agency are morally responsible: 1) moral agency, 2) knowledge, and 3) control. Respectively thus, the entity has to be capable of 1) making normative judgments about its choices and responding correctly to those judgments, 2) obtaining information relevant to this normative assessment, and 3) of being in sufficient control to choose between its options (List, 2021: 16). Which entities meet these conditions is ultimately an empirical matter, List concedes, but he does not see any *a priori* reason to deny moral responsibility to those entities which could be shown to meet them. List considers that the moral agency condition can be met in the form of compliance departments and ethical committees (in the case of corporate agents), rendering it a condition which can be plausibly met by other types of artificial agents as well. This notwithstanding, List is careful to point out that currently feasible artificial agents lack what he takes to be the requisite feature for intrinsic moral significance (phenomenal consciousness) and are thus excluded from the full range of protections and privileges we grant those who have it.

Daniel Tigard (2021) also argues in favor of artificial moral responsibility, but rather than starting from a set of necessary conditions for responsible agency and seeing whether artificial agents can meet them, he takes an ecumenical Strawsonian account of moral responsibility (Shoemaker, 2015) which can accommodate a plurality of agents, and argues that it can be extended to accommodate certain artificial agents as well. According to this account, there are different faces of responsibility – attributability, accountability, and answerability – each of which tracks different agential features – character, regard for others, and evaluative judgments, respectively. Hence, an agent who lacks the feature required for accountability-responsibility (regard for others) might nonetheless be responsible in an attributability or an answerability sense. Tigard suggests that artificial agents with one or more responsibility-relevant feature can thereby qualify as responsible, in that respect.

Diverging from List, who identifies *a priori* what features are required for moral responsibility and then makes a case that artificial agents can meet them, both Tigard and I employ the Strawsonian approach to responsible agency as being a matter of what our practices track. On my analysis, our practices track our capacity to comply with normative demands and, what is more, they cultivate this capacity. Shoemaker's account admittedly offers a richer, more nuanced analysis. However, although his analysis reflects a wider range of psychological and moral commitments, it overestimates the reflective sensitivity of our practices and neglects their regulative power. On my view, our responsibility practices can neither sufficiently reflect nor direct agents' mental content and consequently, they reflect and cultivate only behavioral moral agency. Finally, on the view that Tigard adopts, agents who have particular responsibility deficits – agents on the margins (as Shoemaker puts it) – would not necessarily pass a threshold for responsible agency (should there be one) and would thus have restricted agential status within the moral community. By contrast, my analysis of our practices entails that morally performing agents meet that threshold.

3 THE REGULATIVE NATURE OF MORAL RESPONSIBILITY PRACTICES

The Strawsonian approach takes our practices to reflect responsible agency. Like the moral responsibility consequentialists (Schlick, 1939; Smart, 1961; Dennett, 2015) and instrumentalists (e.g., Vargas, 2013; McGeer, 2019; Jefferson, 2019) and Hume and Hobbes before them,[14] Strawson (2008) acknowledged the regulative power and social utility of our responsibility practices. He simply contended, contra "the optimist" (i.e., the consequentialist), that it would be wrong to account for our practices solely in terms of their effects, as that would undermine their expressive function and their roots in our beliefs – not about regulation – but about desert, responsibility and justice. Additionally, there is the communicative dimension of our practices (Watson, 2004;

[12] By setting strong limits on the degree of punishment, e.g., we administer (eliminating retributivism altogether), attending more, in the ways we can, to agents' reasons, etc.

[13] I hold that robots are desirable only as non-autonomous moral agents, subject to human moral demands since, if morality is a matter of a species' flourishing, as a distinct species, autonomous robots would pursue their own flourishing. Their flourishing may be at odds with ours.

[14] Following thinkers like Thomas Hobbes, Hume points out that rewards and punishments serve to cause people to act in some ways and not in others, which is clearly a matter of considerable social utility (T 2.3.2.5/410; EU 8.2897–98)" (Russell, 2021).

Darwall, 2006; Shoemaker, 2007; McKenna, 2012), according to which they constitute a form of moral address – communicating moral expectations and demands and sustaining interpersonal relationships. However, as the instrumentalists argue (McGeer, 2019; Jefferson, 2019), the regulative effects of these practices are no mere side-effects; our responsiveness to them is constitutive of responsible agency. Furthermore, these practices are necessary for the development and maintenance of responsible agency (Dennett, 2015; McGeer, 2019).

Though I do not deny their expressivist and communicative functions, it is the instrumentalist focus on the role of our practices on moral development that I adopt here. However, whereas the instrumentalist views our practices as necessary and sufficient conditions of robustly responsible agency, I see them as merely necessary. I also claim that they work, in some ways, against the development of moral autonomy. They are necessary because they communicate the normative landscape (Sie, 2018; Sliwa, 2019), regulate behavior in ways that enable internal regulation and reasons-responsiveness, and forge a connection between morally relevant social feedback and behavior. They are insufficient because they cannot enhance moral reasons-responsiveness directly. They are sometimes counterproductive to autonomy because they regulate behavior via conditioning and may impede moral reasons-responsiveness. Briefly, the argument that our practices cannot directly enhance moral reasons-responsiveness goes as follows.[15]

A responsibility response like blame or resentment is surely involved in communicating the normative landscape to developing agents. We can assume, however, that mature wrongdoers, absent excuse, were aware of the relevant moral reason at the time of wrongdoing, in which case the response does not serve to communicate a new moral reason. Although I take it that our responsibility responses are indicative of wrongdoing (we feel resentment, e.g., toward someone who has behaved badly), I hold that they stand at some remove from moral reasons themselves. So with developing agents, resentment (e.g.,) may accompany the moral reason and with both developing and mature agents, resentment may communicate additional moral obligations to the wrongdoer which have been incurred by the wrongdoing – e.g., obligations to express remorse, apologize, reform, etc. However, responsibility responses (reactive attitudes) are not the moral reasons at stake in the wrongdoing and thus can only be paired with moral reasons.

Consider the case of breaking a promise – say to help a friend move, in favor of some selfish motive – say staying on at the sports bar to catch the end of the match. Suppose the motive behind the broken promise comes to light and the promisee resents his friend. The resentment communicates the promisee's disappointment and places the wrongdoer in a position to take further action (expressing remorse, apologizing, promising to uphold promises in the future) should he wish to repair his relationship and moral status. Taking further action manifests regard for the promisee and though the wrongdoer displayed insufficient regard in the initial wrongdoing, I maintain that a general regard for others is insufficient for all our moral obligations (i.e., you can commit a moral wrong while manifesting regard for another's well-being by e.g., lying to spare their feelings). Promise-breaking is wrong irrespective of whether the motive behind the wrongdoing comes to light or the promisee experiences resentment (or even whether a hypothetical agent experiences resentment). Provided thus that our responsibility responses are not themselves the moral reasons at stake in the wrongdoing, and can only accompany moral reasons (or present additional moral reasons), they do not enhance moral reasons-responsiveness directly. Indirect influence cannot guarantee concrete outcomes.

Instead, I suggest, more straightforwardly, that our responsibility responses directly influence only behavior. A behavior cultivation model has a clear evidentiary advantage over the moral reasons-responsiveness cultivation model since we cannot observe agents' mental content directly. On the behavioral model, our practices influence behavior directly by pairing a non-moral reason – the sting or pleasure of the response (e.g., blame or resentment, praise or gratitude) – with the wrong- or right-doing.[16] An agent need only be sensitive to the emotions and opinions of others in order to modify their behavior accordingly. In principle, the higher this sensitivity and the stronger the response, the greater the sting (in the case of blame or resentment) to the wrongdoer, and the stronger a reason to avoid future wrongdoing. In essence, therefore, our responsibility responses require sensitivity, not to moral reasons, but to the pleasure and pain of social approval and disapproval in order to be shaped by them. The very principle of behavioral conditioning is that the reinforced behavior remains after the reinforcing stimulus has been removed. In this respect, we are programming one another,[17] via the moral responsibility practices, to behave according to rules and values rather than to act for the moral reason.

As a brief aside, this description of how our responsibility responses shape behavior may trigger skepticism on the part of the reader as to how non-sentient beings might be responsible. They would not, after all, have the constitution of a human responsible agent – sensitivities to pain and pleasure, approval and disapproval. Though this issue will be addressed in other parts of the paper, a brief clarification is in order. Human moral compliance requires these sensitivities (at least until a feedback independent knowledge of moral reasons and a sensitvity to those reasons arise); machine moral compliance does not. That is not to say that machines need not have "sensitivities" in terms of responsiveness to their programming, but this responsiveness need not resemble ours.

[15] See Gogoshin (2021a) for an elaboration of this argument and the argument in favor of the behavioral model.

[16] Along with Joel Feinberg (1970), I hold that expressions of blame are punishing. I further hold that expressions of praise are rewarding.

[17] I address programming in **Sections 4** and **5**.

Well prior to being able to grasp the moral significance of our actions, we are made, in virtue of these sensitivities, to comply with moral norms. When we are very young, this process is undertaken by our parents and caretakers via the imposition of sanctions and rewards. "Habituation into virtue works because emotional rewards and sanctions gradually alter a person's affective responses and motivational tendencies, in ways that can correct them" (Jacobson, 2005). Once (if) sufficiently habituated to right behavior, we develop an increasingly reasons-responsive disposition and the ability to regulate ourselves. Accordingly, mature agents are not taken to be fitting targets of behavioral management. By a certain level of maturity, educators and caretakers (should) attempt to provide deeper explanations about the moral significance of the actions upon which we impose sanctions and rewards. We hope that over time, children will be motivated by the right/wrong-making features of actions directly. We expect that adults follow laws and moral rules, not out of any fear of getting caught and sanctioned or out of a desire for praise and reward, but out of a deep and well-founded respect for the rightness of those laws and rules (when, of course, those laws and rules are right). We further hope that we will have the capacities to challenge and change those laws and rules which are unjust.

These hopes notwithstanding, by the very nature of behavioral conditioning, as stated, reinforced behavior remains after the reinforcing stimulus has been taken away. Once the connection between action and consequence has been forged, what reason motivates the action – whether the moral reason or the reason tied to the externally imposed (secondary) consequence – may be impossible to discern. On the view that behavior that corresponds with moral norms is moral (or virtuous), this is an unproblematic outcome (from a consequentialist perspective, at least). On the Kantian view, only morally autonomous action – action performed for the moral reason – has moral worth. Any action performed as a result of a law imposed externally (e.g., by means of a sanction) is morally heteronomous (Korsgaard, 1996: 22). However, from an epistemic standpoint, our appraisals of others are generally limited to observables and thus to behavior. We cannot observe reasons for action.

Our very development as moral agents is thus highly dependent, at least early on, on conditioning practices and our means for appraising moral agency, largely limited to appraising behavior. This is not to say that we don't value acting for the relevant moral reason over the prudential one. Our theories of praise and blameworthiness make this distinction; it's our responsibility practices that cannot sufficiently apply it. Furthermore, sanction and reward may well be deeply connected to moral reasons. As previously argued, however, what makes wrong actions wrong and right actions right stands at some remove from sanction and reward and from the reactive attitudes manifested by others. Finally, I suspect that many moral agents develop beyond mere behavioral moral agency. If they do, however, it is likely thanks to something other than what the responsibility system – based on sanction and reward as it is – can provide. Whatever this something consists in, it likely involves institutional support and material conditions with which not all are provided.

4 MECHANISMS OF REGULATION AND THEIR LIMITATIONS

In this section,[18] I address specific features of our responsibility practices which are conducive to a behavioral species of moral agency. First, as conditioning practices, they shape and confine developing agents', in particular, choices. For those who have experienced rewards for certain behaviors will likely be more attentive to these options than those who have not. Still, conditioning does not necessarily bypass the deliberative process. One may contend that anything short of physical coercion shouldn't count as true coercion (Watson, 2004). But the reason for engaging in or avoiding behaviors which have been directly appraised, if the appraisal is effective, might easily become the pursuit or avoidance of these responses, not the moral reason. In fact, our responses may take our attention away from the moral reason, decreasing moral reasons-responsiveness. The fear of social embarrassment alone may easily outweigh a concern for the right reason for one who is not already sufficiently robustly moral reasons-sensitive, and any true wrongdoer is, by definition, insufficiently responsive to moral reasons.

Another agency-defining feature of our practices is their prioritization of behavior over reasons for action and the role this plays in promoting behavioral conformism. As Danaher and Coeckelbergh point out, this prioritization is due in part to our epistemic limitations. In general, we are blind to agents' true motives. It's not to say that we do not care about them and we can of course solicit them from agents post-factum. However, 1) this is generally done only in the case of wrongdoing; we tend not to solicit reasons for right actions, i.e., we tend to take for granted that good-doers have acted for the moral reason and not e.g., to impress their peers.[19] 2) Such testimony is unreliable; we tend to provide post-hoc rationalizations of our behavior (Haidt, 2001), and 3) this is generally relevant only to the way we adjudicate punishment, not to the initial appraisal and response. In general, we attend more to apparent wrongdoing. There is a well-known prioritization of blame (over praise) in our practices (and theories). Moreover, due to 1) above, we often bestow praise upon actions which appear morally worthy even when they're not (e.g., when someone is helpful because they care what by-standers think of them). Even when we don't offer praise, though, by not-blaming these actions, we express approval nonetheless. We thereby promote behavioral conformism, reinforcing behavior which merely conforms with moral values – irrespective of an agent's reasons for acting.

Third, behavioral conditioning via these practices can address only a very limited set of morally-salient behaviors. Insofar as we are wholly dependent on these practices to learn the normative landscape, they can thus provide only limited moral development.

[18] See also Gogoshin (2020).

[19] Though here I make an empirical claim, I take it to be fairly uncontroversial.

1) Their scope is limited to the domain of past actions. We cannot directly influence behaviors which have not occurred. Of course, by letting others know how we will feel or react should they behave in a given way, we can influence their future behavior. a) This would likely be a weaker form of influence than direct, emotional responses and b) the domain of influence is limited to that which can be anticipated and articulated. This form of influence, though part of the inter-personal realm, is akin to the way our society manages our environments, placing limits and negative incentives on certain actions. By including moral reasons and principles of right action along with our responsibility responses, we can target a much wider range of moral behavior. However, provided that we are dependent for right action on these responses (something which is assumed by McGeer, 2019 in her scaffolding view of responsible agency), then our moral agency cultivation remains limited in scope. 2) Acts and expressions of moral condemnation and praise target behavioral outliers – behaviors which transgress or exceed our moral expectations and, of course, only those that are visible to us. On the other hand, not-blaming conformist behavior reinforces it.

Fourth, in order to legitimately hold others to account (e.g., via blame or punishment), we require strong degrees of confidence in their guilt. Although we probe an agent's motivations more deeply in the case of wrongdoing, if we probe far enough below the surface of an agent's history, upbringing, environment, motives, etc., such confidence is hard to come by. Consequently, we tend to base that confidence on seemingly obvious, clear-cut, superficial information about an agent (how the agent appears to us, our perception of their quality of will and motives of action) rather than the deeper but likely truer causal factors at play (see also Dennett, 2015). The result is a restricted set of criteria for our moral responsibility practices which, in turn, fosters a restricted (behavioral) species of agency.

5 THE CASE FOR ROBOT RESPONSIBILITY AND COMMUNITY MEMBERSHIP

As previously stated, on my view (cf. Asaro, 2006), moral agency arises on a spectrum. At the high end of the spectrum is moral autonomy. Somewhere along the spectrum before moral autonomy, the point at which we reach a certain threshold of moral competence, we become morally responsible.[20] I suggest that this competence is the capacity to reliably behave according to moral norms. Without this capacity, we are not morally responsible for our actions and are thereby excluded from the moral community. I argue that moral rule-abiding robots that have the capacity to uphold social role-specific normative expectations are thus morally responsible. According to the Strawsonian notion of moral community as a matter of moral responsibility, morally responsible agents are moral community members too.

Though I leave open the possibility that responsible agency and moral community ought to come apart, there are some normative reasons to keep responsible agency as a sufficient condition of community membership. 1) Responsible agents (agents who can reliably behave in accordance with norms) contribute to the realization of the ethical aim of social cooperation.[21] 2) Demanding more than responsible agency is to demand something our practices cannot (and, in liberal societies, should not attempt to) regulate. Our social responsibility practices regulate behavior (presumably, for the sake of social cooperation). By definition, moral autonomy is not something that can be imposed externally on an agent; it requires that agents be motivated directly by moral reasons. Although there are surely necessary external conditions for the development of moral autonomy (e.g., the right upbringing, a scholarly study of the good,[22] practices which draw our attention to the direct harms and benefits of our actions, thereby cultivating a concern for moral reasons directly rather than for sanctions and rewards), these conditions are not only not guaranteed, they offer no guaranteed outcomes. 3) More troubling, we cannot see or verify whether moral reasons are the motivating reasons. We are largely limited to evaluating and thus enforcing agents' performance.

In support of 1), I offer P. F. Strawson's Strawson (2008: 5) basic conception of morality.[23]

> "Now it is a condition of the existence of any social organization, any human community, that certain expectations on the part of its members should be pretty regularly fulfilled; that some duties, one might say, should be performed, some obligations acknowledged, some rules observed. We might begin by locating the sphere of morality here. It is the sphere of observation of rules, such that the observance of some such set of rules is the condition of the existence of society. This is a minimal interpretation of morality. It represents it as what might literally be called a kind of public convenience: of first importance as a condition of everything that matters, but only as a condition of everything that matters, not as something that matters in itself."

According to Strawson, then, morality in its most basic terms[24] – the observance of a certain set of rules which makes society possible – makes possible the higher human

[20] As a reminder to the reader, by "moral autonomy," I mean governed by (motivated by) the moral reason directly. A morally autonomous agent possesses the capacity to consistently act *for* (not merely in accordance with) the moral reason. This notion is compatible with the Aristotelian ideal of the virtuous person.

[21] I realize that more than behavioral moral agency is necessary for moral progress, for which moral autonomy is necessary (Gogoshin, 2021b).

[22] Following Aristotle in Book II of the *Nicomachean Ethics* (Aristotle and Crisp, 2014).

[23] Thanks to a referee for pointing out two important sources of support for this conception: 1) the morality-as-cooperation view of anthropologist Oliver Scott Curry (Curry et al., 2019) and 2) Joanna Bryson's (Bryson, 2018) view of ethics as being society's means of structuring and maintaining itself, and according to which what is moral is what is socially beneficial.

[24] He acknowledges the inadequacy of this minimal conception of morality, but sees "considerable merit" in it as well.

goods. A moral agent is thus, first and foremost, an agent who follows and whom we expect to follow these rules.[25] Whether a moral agent could or should pursue moral autonomy is irrelevant to their status as a moral agent. Strawson (2008) argues that our moral responsibility responses are reactions to the fulfilling, exceeding, or transgressing of our normative expectations about how others will behave. Moral rule-abiding robots can meet our basic normative expectations and thus support social cooperation.

For humans, meeting these expectations – acting in accordance with moral norms – is no straightforward matter. With robots, again assuming the formalizability and programmability of moral norms, such behavioral compliance is a product of design. This is at odds with a conception of morality as tied to freedom and yet, as previously argued, we are attempting, via the conditioning of the responsibility practices, to program human beings to comply with moral norms too. However, this form of programming can be viewed as a kind of "weak programming" that does not preclude an agent's capacity to alter course. Matheson (2012) argues that sufficiently complex robots can be viewed as weakly programmed as well and so, insofar as humans are weakly programmed and yet morally responsible, so are such robots. As Susan Wolf (1980) has shown, being determined to act morally – as in the case of someone who is incapable of cruelty – is not at odds with moral responsibility; an agent determined in this way is still praiseworthy for their virtuous actions.[26] Finally, to repeat an earlier point, acting against a moral reason and in favor of a selfish impulse is indicative of an agent's lack of moral autonomy.[27] A morally autonomous agent is ultimately responsive to the relevant moral reason. Hence, although I don't consider the robots under consideration in this paper to be morally autonomous, since they are not able to give themselves the moral law (Korsgaard, 1996)[28], their status as programmed entities does not preclude their morally responsible status.

Mature, neurotypical adults are taken to be morally responsible even when they don't behave morally. Moral rule-abiding robots, however, I claim are morally responsible because they have the capacity to reliably behave according to moral norms. As stated previously, it is precisely this capacity that qualifies human agents as responsible agents (see also Dennett, 2015). When a responsible agent transgresses a moral norm, we blame them. However, what renders them liable to blame is their status as a responsible agent, and what gives them this status – machine or flesh and blood – is the capacity to reliably behave according to moral norms. I hold that adults who consistently transgress moral norms, despite being treated as morally responsible, lack this capacity.

At this point, the elephant in the room should be addressed with more than a footnote: whether robots might be capable of acting on moral norms. A significant source of skepticism regarding whether they can rests on the claim that moral agency is a matter of acting for the right reasons which, in turn, requires consciousness (Purves et al., 2015) or the ability to e.g., perceive certain facts as moral reasons (Talbot et al., 2017).[29] Since robots lack these capacities, they lack the relevant capacities for moral agency. On my account, however, responsible agency is a matter of behavior – not mental content. Hence the moral competence of concern to my account is one of performance.

But the elephant remains in the room. Can robots comply with moral norms? And this becomes a matter of whether moral norms can be codified and programmed and then autonomously applied in relevant situations, or whether a design architecture can accommodate learning moral norms from the data and then applying them. Unfortunately, these questions are beyond my expertise to answer; fortunately, they are being addressed.[30] Moreover, there are reasons for optimism on this front, as some autonomous machines (e.g., self-driving cars) are already able to operate relatively reliably in morally and socially significant ways and contexts. They could thus be said to have the moral competence I have argued is relevant to responsible agency. Joanna Bryson's normative argument against the creation of artificial moral agents (see e.g., Bryson, 2018) offers indirect but significant support for the belief that we have or will have the capacity make machines which could behave according to moral norms.

Finally, it is possible that many mature, reasons-responsive agents whom we deem morally responsible are not sufficiently internally regulated or responsive to specifically moral reasons. Our society accordingly manages their behavior by a slightly less visible set of strings – by establishing consequences (largely sanctions) to be imposed by legal and social institutions and by relationship partners (in the form of the negative reactive attitudes if nothing else). Without a reliable means to secure robust responsiveness to moral reasons, it is necessary (and likely more expedient) to rely on our natural aversions to sanction and

[25] See also Gogoshin (2020).

[26] In Dennett's words (Dennett, 2015: 227), "For Kant [. . .] we are only *really* responsible for the right things we do." Wolf provides the contemporary take on it. Like Kant, she does not hold that we are blameworthy for morally wrong actions (though she finds a way to preserve blaming bad behavior). On my view, there is no such asymmetry; however, I do not endorse desert-entailing responsibility. Hence, praise/blameworthy take on a different ring when I use them; i.e., they could stand in for morally right/morally wrong. They could also, taking an instrumentalist or consequentialist rationale, simply denote whether praising/blaming someone can (1) promote their reformation – whether, i.e., they have the right kind of constitution (sensitivities of the sort I have described) to be held morally responsible (Schlick, 1939; Jefferson, 2019) – or (2) be socially beneficial (Dennett 2015; Smart 1961).

[27] This idea, as I understand it, is behind Nadeau's claim (Nadeau, 2006) claim that only androids could be truly moral.

[28] "When you are motivated autonomously, you act on a law that you give to yourself; when you act heteronomously, the law is imposed on you by means of a sanction" (Korsgaard, 1996: 22).

[29] Thanks to the referee who pointed out the need for a clarification here and recommended these references.

[30] See Powers (2006) for a "Kantian machine." See Arkin et al. (2012) for a concrete proposal for moral decision-making in autonomous systems. See Anderson and Anderson (2015) for a principle-based healthcare agent. See Malle and Scheutz (2014) for an environment/feedback moral learning architecture proposal.

desires for reward in order to ensure societal cooperation. Because the moral responsibility practices are regulative and they set, enforce, and reinforce the standards for moral agency and thus moral community membership, they largely both reflect and determine society's level of moral development. Whether this is ultimately desirable is another matter. The point is that the standards at work in our social practices are such that moral rule-abiders qualify as moral community members and what's more, enable social cooperation.

6 SHORTCOMINGS AND POSSIBLE SOLUTIONS

Even should one accept the claim that moral responsibility requires only behavioral moral agency and that some robots can thus be morally responsible, there are moral responsibility functions in terms of accountability which cannot be satisfied by robots. There are of course instances of primitive artifacts (like sex dolls, as noted in Nyholm et al., 2019), not to mention sophisticated androids, which can and will inspire what a human counterpart takes to be a reactive attitude like love. This may not be "genuine love," but even assuming that it is, it's not clear that such a robot could inspire our full range of reactive attitudes. Even if a robot were causally responsible for a mass killing, it's far from certain that we would see any purpose in holding it accountable via blame or punishment. We would, where possible, hold the human moral agents behind the robot morally and criminally responsible. How can we call responsible an agent whom we would not blame, especially if our criteria for responsibility are tied to our practices of holding responsible?

I can offer two answers. 1) As I have argued, having the capacity to reliably behave according to moral norms qualifies one as morally responsible. This claim rests on a distinction made by Angela Smith (Smith, 2007) between the conditions for responsible (blameworthy) agency and the conditions for active blame. The ensuing "gap between conditions of culpability and appropriate blaming, Smith argues, shows that conditions of being responsible cannot be reduced to conditions of appropriate active blaming" (Russell, 2011: 211-212). Hence, robot responsibility is not obviously precluded by the possibility that it might never be appropriate to actively blame them.

2) Our practices are imbued with persistent incompatibilist (libertarian) intuitions. We mistakenly believe that a wrongdoer had the power to do otherwise. Interestingly, responding from this belief to the wrongdoer may be essential for securing forward-looking benefits; i.e., we may have greater success in preventing future wrongdoing when we authentically resent someone for it. Authentic resentment may depend on believing that a wrongdoer ought to have done otherwise. Responding as if the agent really deserves blame or sanction may not only be our only option, psychologically speaking, it may also be the most optimal means of shaping behavior. This said, the very concept of just deserts is the issue at stake in the moral responsibility debate. Would it be fair to blame or punish someone who lacks sufficient control over their character or actions?

The traditional compatibilist says that although we lack ultimate control, we have enough control (or the right kind of control, e.g., guidance control; see Fischer and Ravizza, 2000) to deserve being blamed for our wrongdoings. The instrumentalists, however, have other resources to justify our practices of holding responsible. In conclusion, by rejecting the traditional justification for desert-entailing moral responsibility, we are free to embrace the forward-looking dimension of our practices and hence, the fact that we may not see a backward-looking purpose in holding robots accountable, is not fatal to their moral responsibility. We must thus also consider, which I will do shortly, to what extent robots can be morally responsible in the forward-looking sense.

However, there are other functions served by holding others to account which the above answers do not address. They relate to the inadequate "psychological machinery" (Babushkina, 2020) possessed by foreseeable robots. One such function concerns our primal retaliatory urge, something we share with some primates which, when acted upon, has certain proven physiological benefits for the avenger. This gives rise to what John Danaher (2016) refers to as the "retribution gap."[31] However, revenge practices are also at the root of vicious cycles of aggression and destruction (see Waller, 2012; Waller, 2015). In civilized society, we deter individual acts of revenge and adopt a collective, institutional approach which, though less satisfying to the individual, may nonetheless be characterized as serving a retributive aim. Despite the significant psychological relevance to our practices, as a morally suspect dimension of them (see e.g., Caruso, 2021), I will dismiss this worry as pertains to robots. The issue of adequate psychological machinery is bigger than retributivism, however. Without it, as pointed out by Dina Babushkina (2020), meaningful accountability practices are impossible. Not only can robots not feel the sting of condemnation or punishment or be brought to suffer by them, they cannot feel guilty or, in turn, be forgiven. Blaming robots would thus create a kind of "blame vacuum" and per Danaher (2016) and Babushkina (2020), lead to moral scapegoating.

On the communicative conception, blame is a form of moral address and concerns the blamer and the blamee. Both parties must meet the criteria required for their respective roles. Could a robot meet the criteria for either role? I will focus here on the role of blamee, for it gets to the heart of the concern in the machine moral responsibility debate. According to Coleen Macnamara (2015: 212), eligibility for this role "requires the capacities necessary to give uptake to the distinctive form of communication that reactive attitudes constitute. Uptake of the reactive attitudes amounts to feeling guilt and expressing it via amends, and to respond to blame in this way requires moral competence." Although triggered by a past action, moral address presents forward-looking reasons – apologizing, the making of amends, offering compensation, promising reformation. It is conceivable that robots could fulfill these obligations, at least performatively, but the above objection – when it comes to the psychological dimension of blame – still holds.

[31] Thanks to a referee for providing this reference.

To this objection, I offer the following. 1) On the communicative conception, blame is a two-way street and requires certain symmetrical capacities as relates to communication. Part of the communication is strictly emotional – and the blamer would likely not be psychologically satisfied or able to forgive based on what they take to be a mere performance of guilt or sorrow. Mark Coeckelbergh might respond that robot designers ought to aim for an authentic performance capability and, if successful, it may in fact satisfy the psychological dimension of blame (resolving the "blame vacuum"). If it is not successful, then we face the same problem we presently face with many existing moral community members from whom, when they make moral mistakes, we cannot get the satisfaction or resolution we want from holding to account. This group may include those who have fallen on hard times, those struggling with addiction, mental disorder, poverty, social isolation, poor formative circumstances, toxic social environments, etc. With them, we must establish alternative ways of managing our psychological needs – ways which could then be extended to robots to some degree.

However, if we – as a society – have not reached a legitimate consensus about what societal functions robots ought to fulfill and which robots ought to fulfill them – in light of ultimately transparent and relevant risk-benefit analyses – the blame vacuum is likely to create significant societal problems which I cannot dismiss here. Provided a legitimate consensus, dealing with negative outcomes may be psychologically less challenging than dealing with negative outcomes – even those resulting from strictly human actions – over which we've exercised no agency. 2) Scapegoating is part of the more general responsibility gap problem present in complex technological chains (Matthias, 2004), not just in the context of artificial moral agents. This gap extends well beyond the robot question, through to collective agents and, as I've suggested, to individuals within the human moral community as well.

Where to draw the line for individual human responsibility is no easy task, whether due to determinism or indeterminism in our causal histories. If responsible agency requires a particular psychosocial constitution and careful conditioning practices, which in turn must be tied to the right moral norms, it is likely that many among us are unable to develop responsible agency. For these agents (as well as the abovementioned agents), we need "alternative accountability structures" – social institutions which take responsibility for those who cannot – both the wronged and the wrongdoers.[32] These structures could help equip wrongdoers (or otherwise stand in on their behalf) to fulfill the obligations which their wrongdoing has incurred, in addition to providing them with resources to develop moral competence. These structures should provide recourse to those who have been wronged and who cannot obtain what holding to account should make possible. Such structures could be called upon in the case of robots.

On my account of moral agency, moral autonomy is the level of agency required to be responsible in the deepest sense of the term – to be substantively in control of one's actions and characters[33] – and this level of agency requires practices and conditions on top of our responsibility practices. In this light, only a portion of our moral community is substantively responsible. The majority of the rest of our community has the moral competence to conform to the rules and thus to respond adequately to our practices; the minority does not. The line between these latter two groups, however, is likely very blurry. By minimizing or eradicating traditional desert-based practices and maximizing the forward-looking ones, we reduce the risks of mistaking where this line, if it exists at all, is.[34]

I now return briefly to the forward-looking goals of reformation and restoration independently of any alternative accountability structures. First, robots could be equipped with a primitive reinforcement learning architecture whereby our negative reactions would serve to prevent their negative behaviors in the future (Gogoshin, 2020; Tigard, 2021; see Wallach and Allen, 2009 for an example). Reforming robots directly via the moral responsibility responses is thus conceivable. Moreover, on an instrumentalist account of responsible agency as a matter of susceptibility to our responsibility practices (see Schlick, 1939 for an early version; see Jefferson, 2019 and McGeer, 2019 for more nuanced contemporary versions), robots who could be designed to adequately respond to our practices[35] could thus qualify as fully responsible agents, though not in a way that would satisfy all our folk intuitions and practices. An instrumentalist, however, is well positioned to argue for a revision of our practices. Finally, it is also conceivable that robots may be conscripted to do more extensive (and potentially hazardous) acts of restoration than humans. Hence their forward-looking responsibility may be, in some respects, greater than ours.

The complexity of this discussion, due foremost to the ambivalence which envelops moral responsibility independently of robots, provides an especially weighty reason to reach societal consensus about what roles we want robots to fulfill and what the risk-benefit analysis of having them in these roles amounts to. This would allow us to put the necessary responsibility structures in place such that we can nip at least a bulk of the potential problems in the bud.

[32] This is similar to solutions proposed in response to the responsibility gap (see Behdadi and Munthe, 2020 for a summary). However, I see a responsibility gap even at the level of the human individual. Becoming a responsible individual is itself beyond the control of the individual and sometimes, due to factors beyond society's control as well (e.g., natural misfortunes). Individual responsibility gaps thus abound and society must take responsibility for and within these gaps.

[33] What Hans Jonas (2007) refers to as "substantive responsibility." Compare also Bruce Waller's (Waller, 2012) "take-charge responsibility."

[34] See Gregg Caruso's proposal (Caruso, 2021) for a strictly forward-looking, non-retributivist approach to responsibility and legal justice. He argues that, in the absence of free will, desert-entailing responsibility ought to be rejected. This approach resolves, at least normatively, the particular responsibility (retribution) gap noted in Danaher (2016).

[35] There are many unsettled issues here: what counts as an adequate response, whether the end goal is behavior or full-blown (moral reasons-responsive) moral agency (which it is for McGeer, 2019). But this is a promising path to follow nonetheless for both roboticists and philosophers.

7 CONCLUSION

In this paper, I have argued that the level of moral agency required for moral community membership, insofar as that membership is a matter of responsible agency, is behavioral moral agency. This conclusion is a result of an analysis of the ways our moral responsibility practices function – both in terms of reflecting and fostering moral agency. Given 1) a methodology which takes our practices as evidence of responsibility and the fact that these practices largely address behavior, 2) a conception of morality as a set of rules which enable social cooperation, and 3) the Strawsonian picture of moral community as being a matter of responsible agency, the view that moral rule-abiding robots are responsible and thus moral community members, becomes plausible. Our commitment to moral autonomy necessitates at least two overlapping but distinct conceptions of moral agency. Traditionally, morally responsible agency has been taken to be full-blown moral agency requiring substantive freedom or control, but if our practices are the theoretical starting point, on my analysis of them, this view is incorrect.

I have conceded that robots are unlikely to satisfy all our accountability-responsibility demands. Accordingly, it is vital that we reach the societal consensus described previously. I further proposed what I would propose for the many human moral community members who also lack some degree of accountability-responsibility: alternative accountability structures. Finally, I suggest that we devote resources to the cultivation of human moral autonomy while keeping the bar for moral community membership at responsible agency (as I have defined it herein). This meshes better with our existing moral community, though it also accommodates morally performing agents of any make or model. If this is objectionable, we ought to redefine moral community membership in other terms than morally responsible agency or morally responsible agency in other terms than our responsibility practices.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The author is supported by a personal research grant from the Finnish Cultural Foundation, Grant No. 00190281.

ACKNOWLEDGMENTS

The author thanks the referees for their keen insights and helpful references. She gratefully acknowledges the comradery and excellent scholarship of her fellow RADAR researchers at the University of Helsinki: Raul Hakli, Pekka Mäkelä, Tomi Kokkonen, Pii Telakivi, Olli Niinivara, and Dina Babushkina, whose insights on this and related topics have been invaluable. Thanks are also due to Lilian O'Brien and Antti Kauppinen for their guidance, the audience at Robophilosophy 2020 for helpful comments, and the Finnish Cultural Foundation for financial support. The author is ever grateful to Bruce N. Waller for his expertise and encouragement.

REFERENCES

Anderson, S. L., and Anderson, M. (2015). "Towards a Principle-Based Healthcare Agent," in *Machine Medical Ethics*. Editors S. P. van Rysewyk, and M. Pontier (Cham: Springer International Publishing), 67–77. doi:10.1007/978-3-319-08108-3_5

Aristotle and Crisp, R. (2014). *Nicomachean Ethics*. Revised edition. New York: Cambridge University Press.

Arkin, R. C., Ulam, P., and Wagner, A. R. (2012). Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proc. IEEE* 100, 571–589. doi:10.1109/JPROC.2011.2173265

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. 0 ed. New York, NY: Chapman and Hall/CRC. doi:10.1201/9781420085952

Asaro, P. M. (2006). What Should We Want from a Robot Ethic? *Irie* 6, 9–16. doi:10.29173/irie134

Babushkina, D. (2020). "Robots to Blame," in *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020 Frontiers in Artificial Intelligence and Applications*. Editors M. Norskov, J. Seibt, and O. S. Quick (Washington: IOS Press), 305–315.

Behdadi, D., and Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds Machines* 30, 195–218. doi:10.1007/s11023-020-09525-8

Bryson, J. J. (2018). Patiency Is Not a Virtue: the Design of Intelligent Systems and Systems of Ethics. *Ethics Inf. Technol.* 20, 15–26. doi:10.1007/s10676-018-9448-6

Caruso, G. D. (2021). *Rejecting Retributivism: Free Will, Punishment, and Criminal justice*. Cambridge, United Kingdom; New York, NY: Cambridge University Press.

Champagne, M., and Tonkens, R. (2015). Bridging the Responsibility Gap in Automated Warfare. *Philos. Technol.* 28, 125–137. doi:10.1007/s13347-013-0138-3

Coeckelbergh, M. (2009). Virtual Moral agency, Virtual Moral Responsibility: on the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents. *AI Soc.* 24, 181–189. doi:10.1007/s00146-009-0208-3

Coeckelbergh, M. (2021). How to Use Virtue Ethics for Thinking about the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance. *Int. J. Soc. Robot.* 13, 31–40. doi:10.1007/s12369-020-00707-z

Curry, O. S., Mullins, D. A., and Whitehouse, H. (2019). Is it Good to Cooperate? Testing the Theory of Morality-As-Cooperation in 60 Societies. *Curr. Anthropol.* 60, 47–69. doi:10.1086/701478

Danaher, J. (2016). Robots, Law and the Retribution gap. *Ethics Inf. Technol.* 18, 299–309. doi:10.1007/s10676-016-9403-3

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Sci. Eng. Ethics* 26, 2023–2049. doi:10.1007/s11948-019-00119-x

Darling, K. (2016). "Extending Legal protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects," in *Robot Law* (Cheltenham, United Kingdom: Edward Elgar Publishing), 213–232. doi:10.4337/9781783476732.00017

Darwall, S. L. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, Mass: Harvard University Press.

Dennett, D. C. (2015). *Elbow Room: The Varieties of Free Will worth Wanting*. New edition. Cambridge, Massachusetts; London, England: MIT Press.

Feinberg, J. (1970). *Doing & Deserving; Essays in the Theory of Responsibility*. Princeton, N.J: Princeton University Press.

Fischer, J. M., and Ravizza, M. (2000). *Responsibility and Control: A Theory of Moral Responsibility*. First paperback ed. Cambridge: Cambridge University Press.

Floridi, L., and Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds Machines* 14, 349–379. doi:10.1023/B:MIND.0000035461.63578.9d

Gogoshin, D. L. (2020). "Robots as Ideal Moral Agents Per the Moral Responsibility System," in *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020 Frontiers in Artificial Intelligence and Applications*. Editors M. Norskov, J. Seibt, and O. S. Quick (Washington: IOS Press), 525–534. doi:10.3233/faia200952

Gogoshin, D. L. (2021a). *Taking the reins of moral progress [Presentation]*. In MANCEPT 2021: Moral and Socio-Political Progress, September 8 (University of Manchester).

Gogoshin, D. L. (2021b). *Reactive attitudes and the robustly responsible [Presentation]*. In British Society for Ethical Theory 2021 Graduate Conference, September 17.

Haidt, J. (2001). The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychol. Rev.* 108, 814–834. doi:10.1037/0033-295X.108.4.814

Hakli, R., and Mäkelä, P. (2019). Moral Responsibility of Robots and Hybrid Agents. *Monist* 102, 259–275. doi:10.1093/monist/onz009

Himma, K. E. (2009). Artificial agency, Consciousness, and the Criteria for Moral agency: what Properties Must an Artificial Agent Have to Be a Moral Agent? *Ethics Inf. Technol.* 11, 19–29. doi:10.1007/s10676-008-9167-5

Holroyd, J. (2007). A Communicative Conception of Moral Appraisal. *Ethic Theor. Moral Prac.* 10, 267–278. doi:10.1007/s10677-007-9067-5

Jacobson, D. (2005). Seeing by Feeling: Virtues, Skills, and Moral Perception. *Ethic Theor. Moral Prac.* 8, 387–409. doi:10.1007/s10677-005-8837-1

Jefferson, A. (2019). Instrumentalism about Moral Responsibility Revisited. *Philos. Q.* 69, 555–573. doi:10.1093/pq/pqy062

Johnson, D. G. (2006). Computer Systems: Moral Entities but Not Moral Agents. *Ethics Inf. Technol.* 8, 195–204. doi:10.1007/s10676-006-9111-5

Jonas, H. (2007). *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Chicago: University of Chicago Press.

Korsgaard, C. M. (1996). *Creating the Kingdom of Ends*. Cambridge, New York, NY, USA: Cambridge University Press.

List, C. (2021). Group Agency and Artificial Intelligence. *Philos. Technol.* doi:10.1007/s13347-021-00454-7

Macnamara, C. (2015). "Blame, Communication, and Morally Responsible Agency," in *The Nature of Moral Responsibility: New Essays*. Editors R. K. Clarke, M. McKenna, and A. M. Smith (New York: Oxford University Press), 211–236. doi:10.1093/acprof:oso/9780199998074.003.0010

Malle, B. F., and Scheutz, M. (2014). "Moral Competence in Social Robots," in 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering (Chicago, IL, USA: IEEE), 1–6. doi:10.1109/ETHICS.2014.6893446

Matheson, B. (2012). " Is there a continuity between man and machine?," in *The Machine Question: AI, Ethics and Moral Responsibility* (Birmingham, United Kingdom: The Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB)), 25–28. Available at: http:///www.aisb.org.uk.

Matthias, A. (2004). The Responsibility gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics Inf. Technol.* 6, 175–183. doi:10.1007/s10676-004-3422-1

McGeer, V. (2019). Scaffolding agency: A Proleptic Account of the Reactive Attitudes. *Eur. J. Philos.* 27, 301–323. doi:10.1111/ejop.12408

McKenna, M. (2012). *Conversation & Responsibility*. New York: Oxford University Press.

Nadeau, J. E. (2006). "Only Androids Can Be Ethical," in *Thinking about Android Epistemology*. Editors K. M. Ford, C. N. Glymour, and P. J. Hayes (Menlo Park, CA: Cambridge, Mass: AAAI Press (American Association for Artificial Intelligence); MIT Press, Massachusetts Institute of Technology), 241–248.

Nyholm, S., and Frank, L. E. Philosophy Documentation Center (2019). It Loves Me, it Loves Me Not. *Techné: Res. Philos. Techn.* 23, 402–424. doi:10.5840/techne2019122110

Parthemore, J., and Whitby, B. (2013). What Makes Any Agent A Moral Agent? Reflections on Machine Consciousness and Moral Agency. *Int. J. Mach. Conscious.* 05, 105–129. doi:10.1142/S1793843013500017

Powers, T. M. (2006). Prospects for a Kantian Machine. *IEEE Intell. Syst.* 21, 46–51. doi:10.1109/MIS.2006.77

Purves, D., Jenkins, R., and Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethic Theor. Moral Prac.* 18, 851–872. doi:10.1007/s10677-015-9563-y

Russell, P. (2021). Hume on Free Will. Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/archives/sum2021/entries/hume-freewill/ (Accessed July 15, 2021).

Russell, P. (2011). "Moral Sense and the Foundations of Responsibility," In *The Oxford Handbook of Free Will Oxford Handbooks*. 2nd Edn, Editor K. Robert (Oxford, New York: Oxford University Press), 199–220.

Schlick, M. (1939). *Problems of Ethics*. New York: Prentice-Hall.

Sharkey, A. (2020). Can we program or train robots to be good? *Ethics Inf. Technol.* 22, 283–295. doi:10.1007/s10676-017-9425-5

Shoemaker, D. (2007). Moral Address, Moral Responsibility, and the Boundaries of the Moral Community. *Ethics* 118, 70–108. doi:10.1086/521280

Shoemaker, D. W. (2015). *Responsibility from the Margins*. 1st ed. Oxford, United Kingdom: Oxford University Press.

Sie, M. (2018). "Sharing Responsibility: The Importance of Tokens of Appraisals to Our Moral Practices," in *Social Dimensions Moral Responsibility*. New York, NY, 300–323.

Sliwa, P. (2019). Reverse-engineering Blame 1. *Philos. Perspect.* 33, 200–219. doi:10.1111/phpe.12131

Smart, J. J. C. (1961). I.-Free-will, Praise and Blame. *Mind* LXX, 291–306. doi:10.1093/mind/LXX.279.291

Smith, A. M. (2007). On Being Responsible and Holding Responsible. *J. Ethics* 11, 465–484. doi:10.1007/s10892-005-7989-5

Sparrow, R. (2007). Killer Robots. *J. Appl. Philos.* 24, 62–77. doi:10.1111/j.1468-5930.2007.00346.x

Strawson, P. F. (2008). Social Morality and Individual Ideal. *Philosophy* 36, 1–17. doi:10.1017/S003181910005779X

Strawson, P. F. (2008). *Freedom and Resentment and Other Essays*. London, New York: Routledge.

Sullins, J. P. (2006). When Is a Robot a Moral Agent? *Irie* 6, 23–30. doi:10.29173/irie136

Sullins, J. P. (2011). "When Is a Robot a Moral Agent," in *Machine Ethics*. Editors M. Anderson and S. L. Anderson (Cambridge: Cambridge University Press), 151–161. doi:10.1017/CBO9780511978036.013

Talbot, B., Jenkins, R., and Purves, D. (2017). *When Robots Should Do the Wrong Thing*. Oxford University Press. doi:10.1093/oso/9780190652951.003.0017

Tigard, D. W. (2021). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Camb Q. Healthc. Ethics* 30, 435–447. doi:10.1017/S0963180120000985

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. 1st ed. Oxford: Oxford University Press.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, Mass: Harvard University Press.

Wallach, W., and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford, New York: Oxford University Press.

Waller, B. N. (2012). *Against Moral Responsibility* Boston.

Waller, B. N. (2015). *The Stubborn System of Moral Responsibility*. Cambridge, Massachusetts: MIT Press.

Wallace, R. J. (2011). "Reasons and RecognitionEssays on the Philosophy of T.M. Scanlon," in *Reasons and Recognition: Essays on the Philosophy of T.R. Kumar*. Editors S. Freeman and R. Kumar (Oxford University Press), 307–331. doi:10.1093/acprof:oso/9780199753673.001.0001

Watson, G. (2004). *Agency and Answerability*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199272273.001.0001

Wolf, S. (1980). Asymmetrical Freedom. *J. Philos.* 77, 151. doi:10.2307/2025667

Zimmerman, M. J. (2016). Moral Responsibility and the Moral Community: Is Moral Responsibility Essentially Interpersonal? *J. Ethics* 20 (1–3), 247–263. doi:10.1007/s10892-016-9233-x

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, orclaim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Robotics and AI

BRIEF RESEARCH REPORT
published: 26 November 2021
doi: 10.3389/frobt.2021.788355

Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection

Eric Martínez[1,2]* *and Christoph Winter*[2,3,4]

[1]Massachusetts Institute of Technology, Cambridge, MA, United States, [2]Legal Priorities Project, Cambridge, MA, United States, [3]Instituto Tecnológico Autónomo de México, Mexico City, Mexico, [4]Department of Psychology, Harvard University, Cambridge, MA, United States

To what extent, if any, should the law protect sentient artificial intelligence (that is, AI that can feel pleasure or pain)? Here we surveyed United States adults (n = 1,061) on their views regarding granting 1) general legal protection, 2) legal personhood, and 3) standing to bring forth a lawsuit, with respect to sentient AI and eight other groups: humans in the jurisdiction, humans outside the jurisdiction, corporations, unions, non-human animals, the environment, humans living in the near future, and humans living in the far future. Roughly one-third of participants endorsed granting personhood and standing to sentient AI (assuming its existence) in at least some cases, the lowest of any group surveyed on, and rated the desired level of protection for sentient AI as lower than all groups other than corporations. We further investigated and observed political differences in responses; liberals were more likely to endorse legal protection and personhood for sentient AI than conservatives. Taken together, these results suggest that laypeople are not by-and-large in favor of granting legal protection to AI, and that the ordinary conception of legal status, similar to codified legal doctrine, is not based on a mere capacity to feel pleasure and pain. At the same time, the observed political differences suggest that previous literature regarding political differences in empathy and moral circle expansion apply to artificially intelligent systems and extend partially, though not entirely, to legal consideration, as well.

Keywords: legal personhood, legal standing, moral standing, robot rights, artificial intelligence, artificial intelligence and law, moral circle

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Adam Andreotta,
Curtin University, Australia
Anne Gerdes,
University of Southern Denmark,
Denmark

Correspondence:
Eric Martínez
ericmart@mit.edu

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *02 October 2021*
Accepted: *01 November 2021*
Published: *26 November 2021*

Citation:
Martínez E and Winter C (2021) Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection. Front. Robot. AI 8:788355. doi: 10.3389/frobt.2021.788355

INTRODUCTION

The prospect of sentient artificial intelligence, however distant, has profound implications for the legal system. Moral philosophers have argued that moral consideration to creatures should be based on the ability to feel pleasure and pain (Bentham, 1948; Singer, 1973; Gruen, 2017). Insofar as artificially intelligent systems are able to feel pleasure and pain, this would imply that they would be deserving of moral consideration. Indeed, in their systematic review, Harris and Anthis (2021) find that sentience seems to be one of the most frequently invoked criteria as crucial for determining whether an AI warrants moral consideration. By extension, insofar as the basis for granting legal consideration is based on moral consideration (cf. Bryson, 2012; Bryson et al., 2017), this would further imply that sentient AI would be deserving of protection under the law.

As they stand, however, legal systems by-and-large do not grant legal protection to artificially intelligent systems. On the one hand, this seems intuitive, given that artificially intelligent systems, even the most state-of-the-art ones, do not seem to be capable of feeling pleasure or pain and thus are not eligible for legal consideration (Nevejans, 2016; Bryson et al., 2017; Chesterman, 2020; Andreotta, 2021; but see; Asada, 2019; Shulman and Bostrom, 2021; Galipó et al., 2018). On the other hand, scholars often conclude that artificially intelligent systems with the capacity to feel pleasure and pain will be created, or are at least theoretically possible (Thompson 1965; Aleksander 1996; Blackmore 1999; Buttazzo 2001; Franklin 2003; Harnad 2003; Holland 2007; Chrisley 2008; Seth 2009; Haikonen 2012; Bringsjord et al., 2015; Angel 2019). Furthermore, recent literature suggests that, even assuming the existence of sentient artificially intelligent systems, said systems would not be eligible for basic protection under current legal systems. For example, in a recent survey of over 500 law professors from leading law schools in the United States, just over six percent of participants considered some subset of artificially intelligent beings to count as persons under the law (Martinez and Tobia, 2021).

Moreover, in a separate survey of 500 law professors from around the English-speaking world, just over one-third believed there to be a reasonable legal basis for granting standing to sentient artificial intelligence, assuming its existence (Martinez and Winter 2021a). The study also found that, not only do law professors not believe sentient AI to be eligible for fundamental legal protection under the current legal system, but also that law professors are less normatively in favor of providing general legal protection to sentient AI relative to other neglected groups, such as non-human animals or the environment.

However, it remains an open question to what extent non-experts support the protection of sentient artificial intelligence via the legal system. Surveys of lay attitudes on robots generally suggest that only a minority favor any kind of legal rights in the United States (Spence et al., 2018), Japan, China, and Thailand (Nakada, 2012). Others have found when AI is described as able to feel, people show greater moral consideration (Lee et al., 2019; Nijssen et al., 2019), although it is unclear to what extent this translates to supporting legal protection.

To help fill this void, here we conducted a survey investigating to what extent 1) laypeople believe sentient AI ought to be afforded general legal protection, 2) laypeople believe sentient AI ought to be granted fundamental legal status, such as personhood and standing to bring forth a lawsuit; and 3) laypeople's beliefs regarding legal protection of sentient AI can be accounted for based on political affiliation.

METHODS

Materials

To answer these questions, we constructed a two-part questionnaire, with specific formulations modeled off of recent work by Martinez and Winter (2021a) and Martinez and Tobia (unpublished manuscript).

In the first part (Part I), we designed a set of materials that asked participants to rate how much their legal system 1) descriptively does and 2) normatively should protect the welfare (broadly understood as the rights, interests, and/or well-being) of nine groups:

1) Humans inside the jurisdiction (e.g., citizens or residents of your country)
2) Humans outside the jurisdiction
3) Corporations
4) Unions
5) Non-human animals
6) Environment (e.g., rivers, trees, or nature itself)
7) Sentient artificial intelligence (capable of feeling pleasure and pain, assuming its existence)
8) Humans not yet born but who will exist in the near future (up to 100 years from now)
9) Humans who will only exist in the very distant future (more than 100 years from now)

The two descriptive and normative prompts were presented respectively as follows:

1) One a scale of 0–100, how much does your country's legal system protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?
2) One a scale of 0–100, how much should your country's legal system protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?

With regard to the rating scale, 0 represented "not at all" and 100 represented "as much as possible."

Given that laypeople are not typically experts regarding how the law is or currently works, the purpose of the descriptive question was not meant to establish the ground-truth regarding the inner-workings of the law but rather as a comparison point to the normative question (in other words, to better understand not only how much people think certain groups ought to be protected overall but also how much they think certain groups ought to be protected relative to how much they think they are currently being protected).

In the second part (Part II), we designed materials that related specifically to two fundamental legal concepts: personhood and standing. Personhood, also known as legal personality, refers to "the particular device by which the law creates or recognizes units to which it ascribes certain powers and capacities" (Paton and Derham, 1972; Garner and Black, 1999), whereas standing, also known as *locus standi*, refers to "a party's right to make a legal claim or seek judicial enforcement of a duty or right" (Garner and Black, 1999).

With regard to personhood, we designed a question that asked: "Insofar as the law should protect the rights, interests, and/or well-being of 'persons,' which of the following categories includes at least some 'persons?'" The question asked participants to rate the same groups as in the first part. For each of these groups, the main possible answer choices were "reject," "lean against," "lean towards," and "accept." Participants could also select one of several "other" choices (including "no fact of the matter," "insufficient knowledge," "it depends," "question unclear," or "other").

With regard to standing, we designed a question with the same answer choices and groups as the personhood question but with the following prompt: "Which of the following groups should have the right to bring a lawsuit in at least some possible cases?"

In addition to these main materials, we also designed a political affiliation question that asked: "How do you identify politically?," with "strongly liberal," "moderately liberal," "somewhat liberal," "centrist," "somewhat conservative," "moderately conservative," and "strongly conservative" as the response choices. Finally, we also designed an attention-check question that asked participants to solve a simple multiplication problem.

Participants and Procedure

Participants (n = 1,069) were recruited via the online platform prolific. Participants were selected based on prolific's "representative sample" criteria and were required to be adult residents of the United States.

With regard to procedure, participants were first shown the materials to Part I, followed by the attention check question. Next, on a separate screen participants were shown the materials to Part II. The order of questions in each part was randomized to minimize framing effects.

Participants who completed the study were retained in the analysis if they answered the attention check correctly. Just eight of the original 1,069 participants failed the attention check. We therefore report the results of the remaining 1,061 participants in our analysis below.

Analysis Plan

We analyzed our results using forms of both parameter estimation and hypothesis testing. With regard to the former, for each question we calculated a confidence interval of the mean response using the bias-corrected and accelerated (BCa) bootstrap method based on 5000 replicates of the sample data. In reporting the standing and personhood results, we follow Bourget and Chalmers (2014), Martinez and Tobia (unpublished manuscript), and Martinez and Winter (2021a) by combining all "lean towards" and "accept" responses into an endorsement measure and reporting the resulting percentage endorsement as a proportion of all responses (including "other").

With regard to hypothesis testing, to test whether participants answered questions differently for sentient artificial intelligence relative to other groups, for each question we conducted a mixed-effects regression with 1) response as the outcome variable, 2) group as a fixed-effects predictor (setting artificial intelligence as the reference category, such that the coefficients of the other groups would reveal the degree to which responses for said groups deviated from those of sentient AI), and 3) participant as a random effect.

Because the response scales were different for Parts I and II of the survey, we used a different type of regression model for Parts I and II. For Part I, we used a mixed-effects linear regression. For Part II, we instead used a mixed-effects binary logistic regression, with all "lean towards" and "accept" responses (i.e., those coded as "endorse") coded as a "1", and all other responses (i.e., "lean against," "reject," and "other" responses) coded as a "0."

In order to test the effect of political beliefs on one's responses to the AI-related questions we conducted separate regressions limited to the sentient artificial intelligence responses with 1) response as the outcome variable, 2) politics as a fixed effect (recentered to a −3 to 3 scale, with "centrist" coded as 0, "strongly liberal" coded as 3, and "strongly conservative" coded as -3), and 3) participant as a random-effect.

RESULTS

General Desired Legal Protection of AI

General results of Part I are visualized in **Figure 1**. Of the nine groups surveyed on, sentient artificial intelligence had the lowest perceived current level of legal protection, with a mean rating of 23.78 (95% CI: 22.11–25.32). The group perceived as being most protected by the legal system was corporations (79.70; 95% CI: 78.25–81.11), followed by humans in the jurisdiction (61.88775; 95% CI: 60.56–63.15), unions (50.16; 95% CI: 48.59–51.82), non-human animals (40.75; 95% CI: 39.41–42.24), the environment (40.38; 95% CI: 39.21–41.69), humans living outside the jurisdiction (38.57 (95% CI: 37.08–39.98), humans living in the near future (34.42; 95% CI: 32.83–36.15), and humans living in the far future (24.87; 23.36–26.43).

With regard to desired level of protection, the mean rating for sentient artificial intelligence was 49.95 (95% CI: 48.18–51.90), the second lowest of all groups. Curiously, corporations, the group with the highest perceived current level of protection, had the lowest desired level of protection (48.05; 95% CI: 46.13–49.94). The group with the highest level of desired level of protection was humans in the jurisdiction (93.651; 95% CI: 92.81–94.42), followed by the environment (84.80; 95% CI: 83.66–85.99), non-human animals (73.00; 95% CI: 71.36–74.49), humans living in the near future (70.03; 95% CI: 68.33–71.68), humans outside the jurisdiction (67.75; 95% CI: 66.01–69.42), unions (67.74; 95% CI: 65.96–69.52), and humans living in the far future (63.03; 95% CI: 61.03–64.89).

Our regression analyses revealed the mean normative rating for each group except corporations to be significantly higher than artificial intelligence (p < $2e^{-16}$), while the mean normative rating for corporations was significantly lower than for artificial intelligence (Beta = −2.252, SE = 1.110, p < 0.05). The mean descriptive rating for each group except humans living in the far future was significantly higher than for sentient AI (p < $2e^{-16}$), while the difference between sentient AI and far future humans was not significant (Beta = 1.0132, SE=.8599, p=.239).

When looking at the difference between the desired and current level of protection, seven of the eight other groups had a significantly lower mean ratio between desired and perceived current level of legal protection (p < $8.59e^{-08}$) than artificial intelligence, while the ratios for artificial intelligence and far future humans were not significantly different (p=.685).

With regard to politics, our regression analysis revealed politics to be a significant predictor of participants' response

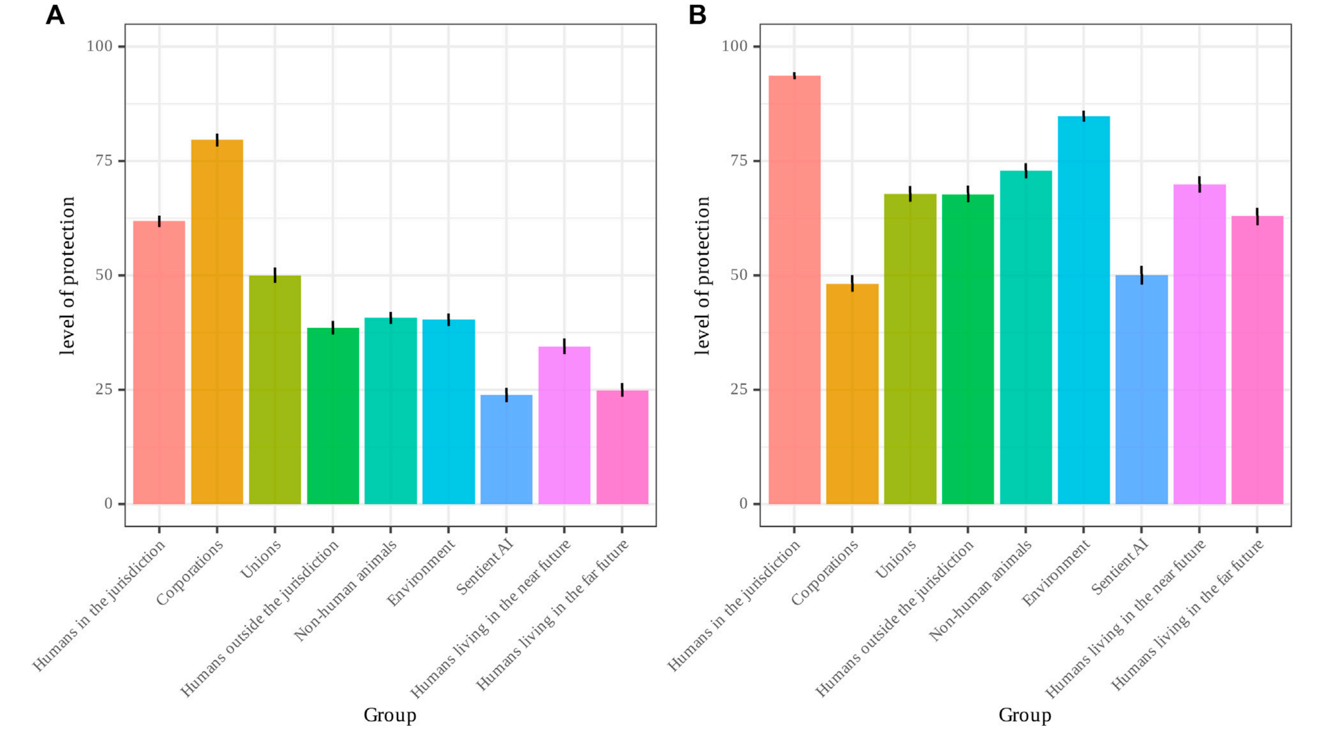

FIGURE 1 | (A) Perceived current and **(B)** desired legal protection for sentient AI and other groups.

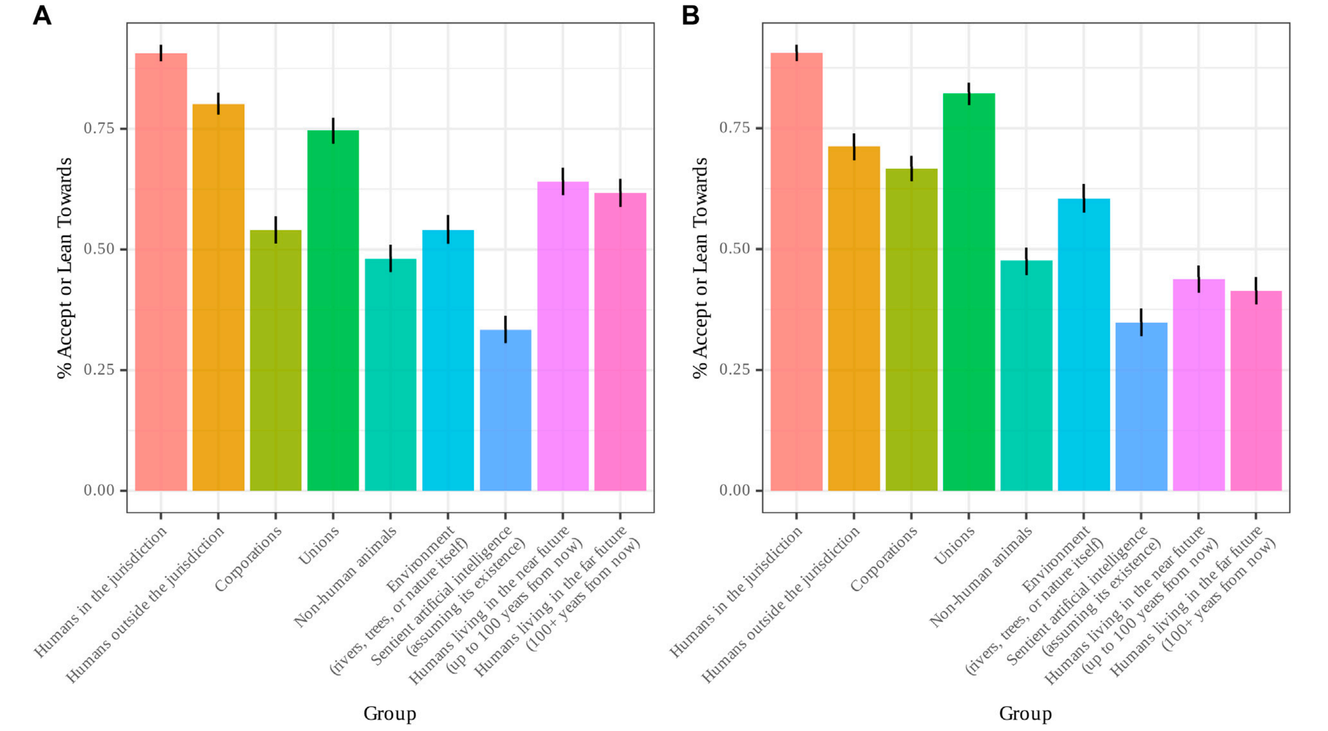

FIGURE 2 | %Endrosement of **(A)** personhood and **(B)** standing for sentient AI and other groups.

to the normative prompt for sentient AI (Beta = 47.9210, SE = 1.1163, p = 1.49e^{-05}), with liberals endorsing a significantly higher desired level of protection for sentient AI than conservatives.

Personhood and Standing

General results of Part II are visualized in **Figure 2**. With regard to personhood, a lower percentage of participants endorsed ("lean towards" or "accept") the proposition that sentient artificial intelligence contained at least some persons (33.39%; 95% CI: 30.71–36.18) than for any of the groups. The next-lowest group was non-human animals (48.12%; 95% CI: 44.87–51.26), the only other group for which less than a majority accepted or leaned towards said proposition. Unsurprisingly, the highest group was humans in the jurisdiction (90.65%; 95% CI: 88.96–92.23), followed by humans outside the jurisdiction (80.16%; 95% CI: 78.10–82.57), unions (74.59%; 95% CI: 71.8–77.21), humans living in the near future (64.09%; 95% CI: 61.33–66.93), humans living in the far future (61.75%; 95% CI: 58.98–64.45), the environment (54.04%; 95% CI: 51.17–57.00), and corporations (53.99%; 95% CI: 51.03–56.86).

With regard to standing, the percentage of participants who endorsed ("lean towards" or "accept") the proposition that sentient artificial intelligence should have the right to bring forth a lawsuit was similarly lower (34.87%; 95% CI: 32.21–37.70) than for all other groups. The next-lowest groups, for whom only a minority of participants endorsed said proposition, were humans living in the far future (41.40%; 95% CI: 38.73–44.33), humans living in the near future (43.80%; 95% CI: 40.72–46.62), and non-human animals (47.68%; 95% CI: 44.73–50.54). The group with the highest endorsement percentage was humans in the jurisdiction (90.60%; 95% CI: 88.89–92.21), followed by unions (82.23%; 95% CI: 79.96–84.50), humans outside the jurisdiction (71.25%; 95% CI: 68.55–73.76), corporations (66.67%; 95% CI: 64.05–69.19), and the environment (60.50%; 95% CI: 57.73–63.54).

Our regression analyses revealed that participants were significantly more likely to endorse personhood (p = 7.42e^{-14}) and standing (p = 1.72e^{-06}) for every other group than sentient AI. With regard to politics, we found a main effect of politics on likelihood to endorse personhood for sentient AI, with liberals significantly more likely to endorse personhood for sentient AI than conservatives (Beta = .098, SE = .036, p=.007). There was no main effect of politics on likelihood to endorse standing for sentient AI (p=.226).

DISCUSSION

In this paper, we first set out to determine people's general views regarding the extent to which sentient AI ought to be afforded protection under the law. The above results paint somewhat of a mixed picture. On the one hand, the fact that people rated the desired level of legal protection for sentient AI as lower than all other groups other than corporations suggests that people do not view legal protection of AI as being as important as other historically neglected groups, such as non-human animals, future generations, or the environment. On the other hand, the fact that 1) the desired level of protection for sentient AI was roughly twice as high as the perceived current level of protection afforded to sentient AI, and 2) the ratio of the desired level of protection to perceived current level of protection was significantly higher for sentient AI than for nearly any other group suggests that people view legal protection of AI as at least somewhat important and perhaps even more neglected than other neglected groups.

The second question we set out to answer related to people's views regarding whether AI ought to be granted fundamental access to the legal system *via* personhood and standing to bring forth a lawsuit. In both cases, the percentage of participants who endorsed the proposition with respect to sentient AI was just over one-third, a figure that in relative terms was lower than any other group surveyed on but in absolute terms represents a non-trivial minority of the populace. Curiously, the endorsement rate among laypeople regarding whether sentient AI should be granted standing in the present study was almost identical to the endorsement rate among law professors in Martinez and Winter (2021a) regarding whether there was a reasonable legal basis for granting standing to sentient AI under existing law, suggesting that lay intuitions regarding whether AI should be able to bring forth a lawsuit align well with legal ability to do so.

On the other hand, the percentage of people who endorse personhood for some subset of sentient AI is several times higher than the percentage of law professors who endorsed personhood for "artificially intelligent beings" in Martinez and Tobia, suggesting either a strong framing effect in how the two surveys were worded or a profound difference in how lawyers and laypeople interpret the concept of personhood. Given that the endorsement percentage for personhood of other groups also strongly differed between the two surveys despite the wording of the two versions being almost identical, the latter explanation seems more plausible. This raises interesting questions regarding the interpretation and application of legal terms and concepts that bear heavy resemblance to ordinary words, as investigated and discussed in previous experimental jurisprudence literature (Sommers, 2020; Tobia, 2020; Martinez and Winter 2021a).

Finally, our study also set out to determine political differences with respect to these questions and found that liberals selected a significantly higher desired level of legal protection for sentient AI and were more likely than conservatives to believe some forms of sentient AI should be considered persons under the law. These findings are consistent with previous literature regarding political differences in moral circle expansion, with liberals tending to display a more universal expanse of empathy and compassion than conservatives (Waytz et al., 2016, 2019). At the same time, the fact that there was no significant difference between liberals and conservatives with regard to standing suggests that the judgment of whether one should have the right to bring forth a lawsuit is not driven by an empathic or compassion-based response to the same degree as in judgments about personhood or general legal protection.

Moreover, liberals and conservatives alike are much less in favor of granting legal protection to sentient artificial intelligence than towards other neglected groups, suggesting that laypeople do not consider the capacity to feel pleasure and pain as sufficient to hold legal rights, similar to the views proposed by scholars that legal personhood ought to be based on autonomy and capacity to act (Solum, 1992; Koops et al., 2010; Laukyte, 2017) or presence and participation in social life (Wojtczak, 2021). Future research could explore to what extent lay attitudes are consistent with these alternative conditions for personhood. Furthermore, given that participants were in favor of increasing legal protection for sentient AI, future research could also explore whether there are other more specific legal rights aside from personhood and standing they might be in favor of so as to satisfy this increased protection.

Although the present study was primarily interested in the descriptive question of to what degree people are in favor of legal protection for sentient AI, one might also attempt to draw normative implications on the basis of our findings. There is a burgeoning literature in the area of experimental jurisprudence dedicated to advancing philosophical, doctrinal and policy arguments on the basis of experimental results (Tobia, 2021a; Sommers, 2021). Within this literature, there is considerable debate as to to what degree and how lay judgments–as opposed to expert judgments–should inform or dictate questions of legal philosophy, doctrine and policy, depending largely on the degree to which one views law through a democratic (as opposed to, say, technocratic) lens (Martínez and Winter, 2021b).

Insofar as one does believe lay attitudes should inform legal doctrine and policy–a view referred to as the folk law thesis (Tobia, 2021b) or the Democratic If-then Approach (Martínez and Winter, 2021b)–the prescriptions one might draw from these results would still potentially remain multifaceted. On the one hand, the fact that laypeople rate the desired level of legal protection to sentient AI as twice as high as the perceived current level, as well as the fact that the difference between the desired and perceived current level of protection was higher than virtually any other group would imply (through this lens) that the existing legal institutions should be reformed so as to increase protection of sentient AI well beyond the current level afforded to them. On the other hand, the fact that the majority of laypeople were not in favor of granting personhood or standing to sentient AI would suggest according to this lens that such increased protection should come in the form of other mechanisms not directly explored in this study, and which, as alluded to before, could be identified through further research projects.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: The datasets for this study can be found on OSF at https://osf.io/2hfx6/?view_only=25d06cdb33004cfa88ac76ae4a28a5b6.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by IRB of Instituto Tecnologico Autonomo de Mexico. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EM and CW contributed to conception and design of the study. EM organized the database and performed the statistical analysis. EM wrote the first draft of the manuscript after discussion with CW. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Legal Priorities Project.

ACKNOWLEDGMENTS

For comments, discussion, and critique, we are grateful to Matthijs Maas, Kevin Tobia, John Bliss and Renan Araújo. We also want to thank Suzanne van Arsdale in particular for help with the manuscript and many useful suggestions.

REFERENCES

Aleksander, I. (1996). *Impossible Minds: My Neurons, My Consciousness*. London, UK: Imperial College Press. doi:10.1142/p023

Andreotta, A. J. (2021). The Hard Problem of AI Rights. *AI Soc.* 36, 19–32. doi:10.1007/s00146-020-00997-x

Angel, L. (2019). *How to Build a Conscious Machine*. New York: Routledge.

Asada, M. (2019). Artificial Pain May Induce Empathy, Morality, and Ethics in the Conscious Mind of Robots. *Philosophies* 4 (3), 38. doi:10.3390/philosophies4030038

Bentham, J. (1948). *An Introduction to the Principles of Morals and Legislation*. New York: Hafner Publishing Co.

Blackmore, S. J. (1999). Meme Machines and Consciousness. *J. Intell. Syst.* 9 (5–6), 355. doi:10.1515/JISYS.1999.9.5-6.355

Bourget, D., and Chalmers, D. J. (2014). What Do Philosophers Believe? *Philos. Stud.* 170, 465–500. doi:10.1007/s11098-013-0259-7

Bringsjord, S., Licato, J., Govindarajulu, N. S., Ghosh, R., and Sen, A. (2015). "Real Robots that Pass Human Tests of Self-Consciousness," in 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN), Kobe, Japan, 31 Aug.-4 Sept. 2015 (IEEE), 498–504. doi:10.1109/ROMAN.2015.7333698

Bryson, J. J. (2012). A Role for Consciousness in Action Selection. *Int. J. Mach. Conscious.* 04 (2), 471–482. doi:10.1142/S1793843012400276

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: The Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25, 273–291. doi:10.1007/s10506-017-9214-9

Buttazzo, G. (2001). Artificial Consciousness: Utopia or Real Possibility? *Computer* 34 (7), 24–30. doi:10.1109/2.933500

Chesterman, S. (2020). Artificial Intelligence and the Limits of Legal Personality. *Int. Comp. L. Q.* 69 (4), 819–844. doi:10.1017/S0020589320000366

Chrisley, R. (2008). Philosophical Foundations of Artificial Consciousness. *Artif. Intelligence Med.* 44 (2), 119–137. doi:10.1016/j.artmed.2008.07.011

Franklin, S. (2003). A Conscious Artifact? *J. Conscious. Stud.* 10 (4–5), 47–66.

Galipó, A., Infantino, I., Maniscalco, U., and Gaglio, S. (2018). "Artificial Pleasure and Pain Antagonism Mechanism in a Social Robot," in *Intelligent Interactive Multimedia Systems and Services*. Editors G. De Pietro, L. Gallo, R. Howlett, and L. Jain (Cham: Springer), 181–189. doi:10.1007/978-3-319-59480-4_19

Garner, B. A., and Black, H. C. (1999). *Black's Law Dictionary*. St. Paul, Minnesota: West Group.

Gordon, J. S. (2021). AI and Law: Ethical, Legal, and Socio-Political Implications. *AI Soc.* 36, 403–404. doi:10.1007/s00146-021-01194-0

Gruen, L. (2017). The Moral Status of Animals, Stanford Encyclopedia of Philosophy. Available at: https://plato.stanford.edu/entries/moral-animal/ (Accessed August 23, 2017).

Haikonen, P. O. (2012). *Consciousness and Robot Sentience. Singapore*. Hackensack, NJ: World Scientific.

Harnad, S. (2003). Can a Machine Be Conscious? How? *J. Conscious. Stud.* 10 (4–5), 69–75.

Harris, J., and Anthis, J. R. (2021). The Moral Consideration of Artificial Entities: A Literature Review. *Sci. Eng. Ethics* 27, 53. doi:10.1007/s11948-021-00331-8

Holland, O. (2007). A Strongly Embodied Approach to Machine Consciousness. *J. Conscious. Stud.* 14 (7), 97–110. doi:10.1016/j.neunet.2013.03.011

Koops, B., Hildebrandt, M., and Jaquet-Chiffelle, D. (2010). Bridging the Accountability Gap: Rights for New Entities in the Information Society? *Minn. J. L. Sci. Technol.* 11 (2), 497–561.

Laukyte, M. (2017). Artificial Agents Among Us: Should We Recognize Them as Agents Proper? *Ethics Inf. Technol.* 19, 1–17. doi:10.1007/s10676-016-9411-3

Lee, M., Lucas, G., Mell, J., Johnson, E., and Gratch, J. (2019). "What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations," in *Proceedings of the 19th Acm International Conference on Intelligent Virtual Agents*, 38–45.

Martinez, E., and Tobia, K. P. (2021). *The Legal Academy and Theory Survey*. (unpublished manuscript).

Martínez, E., and Winter, C. (2021a). *Protecting Future Generations: A Global Survey of Legal Academics*.

Martínez, E., and Winter, C. (2021b). *Advances in Experimental Philosophy of Law (Forthcoming), Legal Priorities Project Working Paper Series*. Experimental Longtermist Jurisprudence (2).

Nakada, M. (2012). *Discussions on Robots and Privacy as Topics of Intercultural Information Ethics in 'Far East'*. Robots and Privacy in Japanese, Thai and Chinese Cultures.

Nevejans, N. (2016). European Civil Law Rules in Robotics. Directorate-General for Internal Policies, Policy Department C: Citizens' Rights and Constitutional Affairs. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf (Accessed October 31, 2021).

Nijssen, S. R. R., Müller, B. C. N., Baaren, R. B. V., and Paulus, M. (2019). Saving the Robot or the Human? Robots Who Feel Deserve Moral Care. *Soc. Cogn.* 37 (1), 41–S2. doi:10.1521/soco.2019.37.1.41

Paton, G. W., and Derham, D. P. (1972). *A Textbook of Jurisprudence*. Oxford: Clarendon Press.

Seth, A. (2009). The Strength of Weak Artificial Consciousness. *Int. J. Mach. Conscious.* 01 (01), 71–82. doi:10.1142/S1793843009000086

Shulman, C., and Bostrom, N. (2021). "Sharing the World with Digital Minds," in *Rethinking Moral Status*. Editors S. Clarke, H. Zohny, and J. Savulescu (London: Oxford University Press), 306–326. doi:10.1093/oso/9780192894076.003.0018

Singer, P. (1973). "Animal Liberation," in *Animal Rights*. Editor R. Garner (London: Palgrave Macmillan), 7–18. doi:10.1007/978-1-349-25176-6_1

Solum, L. B. (1992). Legal Personhood for Artificial Intelligences. *North Carolina L. Rev.* 70 (4), 1231–1287.

Sommers, T. (2020). Commonsense Consent. *Yale L. J.* 129 (8), 2232–2324. doi:10.2139/ssrn.2761801

Sommers, R. (2021). Experimental Jurisprudence. *Science* 373 (6553), 394–395. doi:10.1126/science.abf0711

Spence, P. R., Edwards, A., and Edwards, C. (2018). "Attitudes, Prior Interaction, and Petitioner Credibility Predict Support for Considering the Rights of Robots," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 243–244. doi:10.1145/3173386.3177071

Thompson, D. (1965). Can a Machine Be Conscious? *Br. J. Philos. Sci.* 16 (61), 33–43. doi:10.1093/bjps/xvi.61.33

Tobia, K. P. (2020). Testing Ordinary Meaning. *Harv. L. Rev.* 134 (2), 726–806. doi:10.2139/ssrn.3266082

Tobia, K. (2021a). *Experimental Jurisprudence. 89*. University of Chicago Law Review (2022, Forthcoming). doi:10.2139/ssrn.3680107

Tobia, K. (2021b). *Law and the Cognitive Science of Ordinary Concepts. Law and Mind: A Survey of Law and the Cognitive Sciences*.

Waytz, A., Iyer, R., Young, L., and Graham, J. (2016). "Ideological Differences in the Expanse of Empathy," in *Social Psychology of Political Polarization*. Editors P. Valdesolo and J. Graham (New York: Routledge), 61–77. doi:10.4324/9781315644387-4

Waytz, A., Iyer, R., Young, L., Haidt, J., and Graham, J. (2019). Ideological Differences in the Expanse of the Moral Circle. *Nat. Commun.* 10 (1), 1–12. doi:10.1038/s41467-019-12227-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 02 December 2021
doi: 10.3389/frobt.2021.769349

Speculating About Robot Moral Standing: On the Constitution of Social Robots as Objects of Governance

*Jesse De Pagter**

Institute for Management Science, TU Wien, Vienna, Austria

In recent years, the governance of robotic technologies has become an important topic in policy-making contexts. The many potential applications and roles of robots in combination with steady advances in their uptake within society are expected to cause various unprecedented issues, which in many cases will increase the demand for new policy measures. One of the major issues is the way in which societies will address potential changes in the moral and legal status of autonomous social robots. Robot standing is an important concept that aims to understand and elaborate on such changes in robots' status. This paper explores the concept of robot standing as a useful idea that can assist in the anticipatory governance of social robots. However, at the same time, the concept necessarily involves forms of speculative thinking, as it is anticipating a future that has not yet fully arrived. This paper elaborates on how such speculative engagement with the potential of technology represents an important point of discussion in the critical study of technology more generally. The paper then situates social robotics in the context of anticipatory technology governance by emphasizing the idea that robots are currently in the process of becoming constituted as objects of governance. Subsequently, it explains how specifically a speculative concept like robot standing can be of value in this process.

Keywords: anticipatory governance, object of governance, robot ethics, robot governance, robot standing, speculative concept

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Joshua Gellers,
University of North Florida,
United States
Maciej Musiał,
Adam Mickiewicz University, Poland

****Correspondence:***
Jesse De Pagter
jesse.de.pagter@tuwien.ac.at

Specialty section:
This article was submitted to
Ethics in Robotics and Artificial
Intelligence,
a section of the journal
Frontiers in Robotics and AI

Received: *01 September 2021*
Accepted: *03 November 2021*
Published: *02 December 2021*

Citation:
De Pagter J (2021) Speculating About Robot Moral Standing: On the Constitution of Social Robots as Objects of Governance.
Front. Robot. AI 8:769349.
doi: 10.3389/frobt.2021.769349

INTRODUCTION

In recent years, the governance of robotic technologies has become an increasingly prominent issue within policy-making contexts. An important motivation behind the proclaimed need for such governance is that anticipatory approaches are crucial in order to keep pace with imminent transitions within society as the implementation of robots becomes increasingly widespread (Taeihagh, 2021). Such concerns over the insertion of robots into existing social contexts can at least partly be explained with reference to the widely diverging, speculative trajectories connected to the future of (social) robots (Suchman, 2019). These include predictions concerning the increasing applications and roles of (humanoid) social robots, which could potentially pose crucial challenges to the way social life has been organized for many years (Kim and Kim, 2013). Within the discussion on those challenges, the notion of robot standing is currently an increasingly important yet controversial concept. Complicating this discussion is the fact that in many cases, what needs to be governed - the widespread implementation of robots that could bring about fundamental societal transformations - has not yet been realized. While there are many signs and signals that such robots will or could soon be implemented on a broad scale, they are mostly currently still in investment and development stages (Mindell, 2015). Questions and debates regarding social life with robots therefore have quite a speculative character, and their relevance is sometimes questioned. The discussion on robot standing can be seen as an illustrative case of a controversy that is heavily vested in forms of speculative anticipation about the future of robots. This paper will take a closer look at the speculative character of the robot standing concept and discuss its usefulness for the process of constituting social robots as objects of governance.

The robot standing concept posits that artificial agents could have claims to novel forms of moral and/or legal status (Coeckelbergh, 2014). Thus, it is closely related to discussions on new understandings of technological artifacts and related changes in the conceptualization of agency. This paper does not provide new conceptualizations or ideas related to the discussion on robot standing itself; rather, it reflects on the usefulness of having such a discussion. That is to say, the speculative content of the robot standing concept is argued to be instrumental for the process of constituting robots as objects of governance. This process should be understood as open-ended: as (many types of) social robots are still emerging as technological artifacts of which the implementation has not yet fully materialized, so are the conceptual schemes that need to be developed to interpret and deal with the societal implications of those robots. The goal of this paper is first of all to provide new directions in the discussion of the significance and usefulness of speculative concepts like robot standing, by arguing that it can guide the development of ideas behind anticipatory robotic governance. In the context of fastly emerging robotics and AI, anticipatory governance is currently a prominent issue, as the main objective of such governance is to manage emerging technologies, while such management is still possible (Guston, 2014; Winfield and Jirotka, 2018) Second, in so doing, the paper is also meant to provide new arguments in response to opponents of the very idea of robot standing, who deem it irrelevant or harmful (e.g. Bryson, 2010; Birhane and van Dijk, 2020; Pasquale, 2020).

Therefore, while the debate on robot standing can be understood as an example of explicitly speculative engagement with emerging technology, this paper argues that speculative thinking on moral standing is an important and fruitful part of the process of robots becoming objects of governance. It does so via the following structure: The section below introduces important concepts underlying the notion of robot standing while summarizing the arguments of several voices in the debate. Based on this discussion, the section then examines the speculative elements inherent in the robot standing concept while also outlining the wider debate on the role of speculative concepts within the critical study of technology. The next section discusses what it means to understand robots as objects of governance. It explains how the process of constituting an object of governance should be understood, thereby elaborating on the role of speculative concepts like robot standing. The section after that discusses how the concept of robot standing itself can play a role in such a process when it comes to robotic governance. Finally, the conclusion will provide a short reflection on the role of philosophy of technology in the development of speculative concepts.

SPECULATIVE ELEMENTS IN THE DISCUSSION ON ROBOT STANDING

Concepts with speculative content can be helpful to anticipate technological potential. Nevertheless, the analysis of unrealized technological potential is an ambivalent topic in contemporary philosophy of technology as well as in other (social constructivist) fields that analyze the relationship between technology and society. In principle, the idea of social robots' potential is rather straightforward, namely that the many different robotic technologies currently under development are accompanied by different expectations and promises regarding the future possibilities that those technologies present. However, as already indicated above, even though some new types of social robot technology might already be reality, many anticipated robots are still in the research and development process. At the same time, the public is teased with demonstrations of social robots, which are nevertheless largely still not part of daily social life. Autonomous social robots can therefore be understood to be in a phase where their sociotechnical potential is still mostly unrealized, while their implementation is simultaneously very much anticipated. Within the academic fields engaged in the critical study of technology, it is rather common to be very critical of such signs and signals of new futures. Moreover, conceptualizing the notion of future potential has proven to be difficult, especially when trying to abstain from determinist or instrumentalist views of technology, both of which are often seen as problematic (Wyatt, 2008; Dafoe, 2015). In fact, it is common practice in philosophy and the (qualitative) social sciences to analyze and often even debunk speculations regarding technological futures as a form of hubris. Technological potential, in such cases, is often implicitly or explicitly assumed to be conceptually problematic, theoretically incomprehensible, or denounced as a deterministic element in the discourse surrounding the technology under study (Heilbroner, 1994; Cressman, 2020). However, a possible way to engage with technological futures is to anticipate them by engaging with them while trying to analyze the ramifications of certain specific potentialities. I argue here that the debate on robot standing occupies an interesting position in this regard, as its engagement with the future potential of robotic technology contains elements that are explicitly speculative. As such, it is currently a relevant yet controversial concept that has already invited many different thinkers to engage with the possible consequences of robots as artificial agents.

Before delving into the topic of robot standing and its speculative character, it is useful to provide a short definition of what the notion of a "speculative concept" means in this context, especially since the term "speculative" has many different connotations. Speculative concepts, in this specific framework of emerging technology and its governance, can first of all be defined as concepts that aim to engage with the sociotechnical potential of an emerging technology. Sociotechnical potential in this case simply means that a multifaceted network of social and technical elements is considered during the assessment of that technology's societal impact (Cressman, 2020). Furthermore, from this perspective, the sociotechnical potential of a specific technology is explicitly understood to be in a continuous state of controversy due to its undetermined character. Second, speculative concepts are understood to assist in the delineation of anticipatory scenarios based on actual developments. How realistic such anticipatory scenarios are, however, is always up for discussion, especially because engagement with the possible futures of technology already implies specific types of unknowns and contingencies. Third, the emphasis on speculative concepts as *concepts* is crucial. Concepts can be applied, discussed and reconceptualized in different contexts and can change their meaning depending on them. Finally, concepts are also different from overarching philosophical theories. Speculative concepts are in that regard smaller entities than theories. While they can certainly draw inspiration from larger philosophical frameworks, they are usually easier to apply in settings outside of these philosophical traditions.

Introducing the Robot Standing Issue

From a broader philosophical point of view, the notion of robot standing and the arguments surrounding it can be seen as part of a general cultural fascination with machines as lively beings - a fascination which includes frequently mentioned historical examples such as Henri Maillardet's automaton or Japanese karakuri puppets (Rossi et al., 2009). These examples of automata demonstrate how the notion of robots having a certain kind of standing, be it social, moral or otherwise, is part of the human fascination with alternative (non-human) forms of agency (Lindstrøm, 2015; Heffernan, 2019). However, this is not an easy topic, since objects, in whatever form, have been quite systematically barred from having any form of agency in modern societies (Harman, 2016). Generally, recent decades have seen rising interest in new forms of ontological pluralism and ethical extensionism, which pose novel ways of looking at objects in general and technological artifacts in particular (Chan, 2011; Pickering, 2017). As a part of this development, many different theories of non-human forms of agency have been developed. Bruno Latour, for instance, famously argued that modernity's traditional subject-oriented moral theories conceal the agency and demands of non-human entities (Latour, 2005, 2014). In recent decades, several different academic fields, mainly in the social sciences and humanities, have either developed materialist critiques based on ideas of non-human agency, or have at least derived inspiration from those ideas (Law, 2008). Often, these theories and methods explicitly understand artifacts to carry forms of inherent sociality while emphasizing the (moral) agency of non-human entities like, for instance, technological artifacts (Gunkel, 2012). Many of those theories have speculative content or are based on concepts and ideas that are explicitly speculative in the sense that they refer to potential futures with new forms of agency. Others are based on entities that do not yet exist but can be anticipated. An important example of a speculative notion that is often mentioned in this context is Donna Haraway's concept of the cyborg, which is developed to explore its emancipatory potential and unsettle solidified societal assumptions (Haraway, 1991).

Theorists like Latour and Haraway have conducted groundbreaking work on fundamentally novel ways of understanding and theorizing social agency. Although their theories and ideas do not explicitly engage with the topic of robot standing and its ramifications, an important discussion related to their endeavours is that of the human-machine boundary (Suchman, 2006). This discussion has become increasingly prominent in various academic fields during the last decade, as new developments in autonomous technology sparked an interest in exploring the implications and complications of such technologies (Floridi and Sanders, 2004; Dautenhahn, 2007). If they were to become reality on a wide scale, autonomous social robots are set to disturb modernist understandings of fundamental notions that are integral to the boundary between humans and machines, such as (moral) agency, responsibility, personhood, or empathy (Wallach and Allen, 2009). Several of those basic concepts are considered to be important to human identity and as such, have played a critical role in many (Western) legal, psychological and social concepts (Koops et al., 2013; Alač, 2016; Danaher, 2019; Fosch-Villaronga et al., 2020). If (social) robots were indeed to disturb such concepts, this could have profound implications for how humans understand themselves and how their societies are organized (Sætra, 2021). In that regard it is useful and important to think about the ethics of non-human entities (Gellers, 2020). For instance, synthetic persons, under which social robots would fall, present significant legal lacunae when it comes to most countries' current legal systems (Bryson et al., 2017). Whereas most voices in this discussion would probably hesitate to ascribe proper sentience to robots, an important argument in the debate on standing is the discussion on the agentic appearance of social robots and the agency that should be attributed on the basis of that (Coeckelbergh, 2010; Nyholm, 2018). In this regard, the future potential of social robots becoming perceived as autonomous agents is generally an important topic in robotics research. There is already a lot of research in more applied fields like Human Robot Interaction (HRI) anticipating the agentic appearance of robots by applying so-called "Wizard of Oz studies", in which robots' autonomy and agency is imitated in order to conduct research about how humans would react to robots' appearances and actions if they were to have agentic qualities (Maulsby et al., 1993; Riek, 2012). Closely related to this issue of appearance is the issue of control: Autonomous, agentic action by a machine assumes a certain lack of control by humans (Coeckelbergh, 2015; Wallach, 2015). This is also where important questions arise with respect to the governance of machine agency and the concepts of moral and legal standing attached to it. Hence, it is no surprise that (social) robots are being studied by legal philosophers and ethicists. Indeed, the regulation of robots, often in combination with artificial intelligence, has become an important topic in this field in recent years (Pagallo, 2013; Leenes et al., 2017; Turner, 2019). David Gunkel, an important proponent of the discussion on robot rights, nicely summarizes this by writing that "the question of robot rights (assuming that it is desirable to retain this particular vocabulary) makes a fundamental claim on ethics, requiring us to rethink the systems of moral considerability all the way down" (Gunkel, 2018a, 185).

Furthermore, the issue of robot standing has recently also started to become an actual topic in policy-making. An important example that is often mentioned in this context is the European Parliament (EP) considering the idea of electronic personality. This is not necessarily the same as robot standing, but certainly bears similarities in terms of its underlying dynamic. The EP's report suggests the following with respect to the legal and economic notion of "electronic personality" (EP, 2017, §59f):

> "creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently"

In this quote, the EP argues that the actions and responsibilities of robots will render electronic personhood necessary in order to deal with their economic and legal consequences. Implicit in this understanding of such personhood is a notion of robot standing based on responsibility. It is exactly within such a context that attempts at anticipatory governance can be seen as guided by speculative concepts like robot standing. Nevertheless, this EP proposal immediately exposes the controversy of the issue, as it received serious backlash: an open letter was signed by 156 artificial-intelligence experts from 14 European countries, rejecting the EP's recommendations (Nevejans, 2018). Thus, the fact that the autonomy of robots engaged in different types of social interactions could lead to significant challenges to the basic underpinnings of societal and legal understandings does certainly not mean that the participants in the debate are agreeing on robot standing. In fact, many consider the question of robot standing and the related idea of robot rights, very problematic. For instance, Joanna Bryson writes that there can be no real discussion about rights, since in the end robots are owned by humans (Bryson, 2010). Others call for a shift in focus towards safeguarding the welfare of all humans rather than focusing on robots while denouncing the issue of robot rights as something for AI and robotics futurists (Birhane and van Dijk, 2020; Pasquale, 2020). Furthermore, legal scholars have explicitly argued that robots should be deemed products, thereby excluding any considerations that understand robots as bearers of any rights or obligations (Bertolini and Aiello, 2018). Keeping this in mind, the goal of this paper is not necessarily to take a strong side in those debates, but much rather to explicitly consider the role of speculative content implicit in the robot standing concept and reflect on it as such. In order to do that, we must take a step back and be more explicit about the character of this speculative content, which will be done below.

Robot Standing and Its Speculative Ethics

Whereas the arguments above demonstrate various ideas about robot standing, it is important to seriously consider whether the discussion as a whole is too far-fetched and excessively rooted in speculative, futuristic arguments that bear no ground in engineering reality. David Gunkel, who was already mentioned above, takes an important, quite distinctive voice in this debate, as he strongly argues for exploring "robot rights", an issue closely connected to the topic of robot standing (Gunkel, 2018a; 2018b). In his book on robot rights, he explains and reviews the different positions on the question of robot standing. Gunkel quotes and refers to an array of philosophers who are mostly sceptical about the usefulness of the notion of robot rights, a notion that is closely related conceptually to robot standing. The main point in this view is that robot rights are "unthinkable". Gunkel himself counters this criticism by arguing that it is a task of critical thinking to expose why the unthinkable is unthinkable, thereby "confronting and thinking the unthinkable" (2018a, 51). Furthermore, he argues that ethics is the field with the tools and obligations to ultimately challenge the status quo, which is exactly how moral theories and practices evolve. The task of ethics, he writes, is to "stress-test and question the limitations and exclusions of existing moral positions and modes of thinking. Defending orthodoxy is the purview of religion and ideology; critically testing hypotheses and remaining open to revising the way we think about the world in the face of new challenges and opportunities is the task of science" (2018a, 52).

Gunkel's focus on the role of ethics is interesting here, as the field generally has a rather unique position when it comes to engagement with speculative technological futures. Much of the philosophical work focused on ethical thinking with regards to technological development is in fact participating in the anticipation of future social and legal ontologies. That is to say, ethicists who study robotics (or other emerging technologies, e.g. nanoethics) often actively engage with questions that are somewhat speculative in order to discuss ethical challenges and lacunae that the future of those technologies could bring about. One might think this only applies to posthumanist ethics, but this is certainly not the case. Many of the current discussions around social robots in philosophy are focused on describing and analysing new ontologies regarding the human-machine boundary. Accordingly, ethicists have extensively engaged in speculative explorations of future legal and social ontologies and their consequences for human social life with robots. Within philosophy, the examination of such questions and their potential implications has been a natural fit for several of its subdisciplines, presenting a great opportunity to gain practical and effective relevance in a society that is increasingly organized around expertise. Furthermore, this type of engagement has arguably increased ethicists' interdisciplinary collaboration with many other fields in robotics, such as HRI, legal theorists, robot engineers and so on.

The question remains why such interdisciplinary ethics approaches based on speculative concepts are considered problematic. One of the staunchest critics of this speculative element in the ethics of emerging technology, Alfred Nordmann, provides clear insight into this issue. In his ethical and technophilosophical deliberations on the future of nanotechnology, Nordmann strongly argues against what he calls "speculative nanoethics" which, he argues, is based on the technological hubris of "if-and-then" rhetorics (Nordmann, 2007). Nordmann, who refers to himself as a "reluctant ethicist", problematizes various ethical approaches that are imaginative with respect to the future, exhibiting a clear preference for less imaginative approaches that "bring to light how less spectacular, more familiar technologies shape and reshape, perhaps transform social interactions, individual agency, and a sense of subjectivity or self" (p. 44). In a paper with Arie Rip, Nordmann writes that "worries about the most futuristic visions of nanotechnology can cast a shadow on all ongoing work in nanoscience and technology" (Nordmann and Rip, 2009, 274). By making these points, Nordmann started a fruitful and important discussion within the field of nanoethics, but also in the larger context of the critical analysis of anticipatory approaches (Nordmann, 2014). Various other works have since discussed arguments complementary to Nordmann's. For instance, Ibo van de Poel proposes an alternative to speculative anticipatory approaches when he argues for the gradual experimental introduction of new technologies, while assessments regarding the acceptability of such introductions should be based on ethical frameworks (van de Poel, 2016). On the other hand, Nordmann's arguments have also been strongly criticized. For example, Armin Grunwald argues that instead of 'speculative ethics', we should speak about 'explorative philosophy' which "must develop methods and procedures of assessing pictures of uncertain futures with respect to their degree of rationality" (Grunwald, 2010, 99). Cynthia Selin writes in a direct response to (Nordmann, 2014) article that "foresight practices are meant to contrast the techno-scientific, future-grasping hubris that has been under scrutiny from STS scholars (amongst others) for decades," while also writing that Nordmann fails to systematically categorize what forms of speculation exactly are unacceptable (Selin, 2014, 103).

Whereas this discussion on the role of speculative concepts has mostly been confined to insiders within academic fields such as philosophy of technology and science and technology studies (STS), the notion of robot standing and its speculative character have caused a stir both inside and outside of academia. As such, it is a particularly good example of the role of speculative concepts in the analysis of (emerging) technology. In this context, it is interesting when David Gunkel writes that "science fiction is both a useful tool for and a significant obstacle to understanding what the term "robot" designates" (2018a, 18). Importantly, Gunkel emphasizes here the importance of understanding that what "robot" means is socially negotiated and that "word usage and terminological definitions shift along with expectations for, experience with, and use of the technology" (2018a, 23). Those quotes already provide an indication on how speculative concepts like robot standing can be useful from an anticipatory governance perspective. First of all, it is particularly challenging to engage in anticipatory governance that prepares for futures involving potentially disruptive technologies. While it has already been demonstrated that the development and application of speculative concepts is a contested practice in general, my goal is to further establish the development and implementation of specific kinds of speculative thinking *within* the empirical tradition of the critical study of technology. This research tradition has already provided very relevant insights for policy ideas while directly engaging with technology in the making via both philosophical and (qualitative) social science methods (see e.g. Boden et al., 2017; Bösl and Bode, 2018; AIHLEG, 2019). Robotic technologies represent a great example of this type of engagement since their societal impact is currently highly anticipated. Furthermore, as will be argued below, the concept of robot standing provides valuable insight into the way a speculative concept can be used in the (empirical) critical study of technology and its governance challenges. Even if one agrees with (some of) the problematizations concerning Nordmann's so-called "if-and-then" rhetoric, the main point here is that it remains important to engage with the issue of future contingency in technological development and its governance through concepts like robot standing and the debates around it. It is exactly in such a context that the robot moral standing concept is explored in the following section.

ROBOTS AS OBJECTS OF FUTURE GOVERNANCE

The main point of this section is to argue how a speculative concept like robot standing can be of value in the process of constituting robots as objects of governance. This process is explicitly understood to be far from completed, and the goal is to develop an argument that explores speculative thinking on moral standing as an important and worthwhile element of this process. It should be mentioned in this regard that in several policy areas, robots are already very much constituted as objects of governance. For instance, industrial robots have been used in industry for many years. In this context, policies regulating and governing robots are clearly established, such as in terms of safety and liability: for instance in the context of the EU, very specific rules apply when it comes to safety and industrial robots, regulated by policies such as the Machinery Directive (Directive 2006/42/EC), the Framework Directive for Occupational Safety and Health (Directive 89/391/EEC) and others, often depending on the context of use. In this case, robots are mostly defined (and thus also regulated) as being possibly dangerous to workers' health and safety. Furthermore, robots have long been a part of policy discourse in strategic economic policy-making, in which their presence has unsurprisingly become an indicator of an economy's rate of automation, innovation and economic progress. However, the main issue in the case of the discussion around robot autonomy is not how robots are currently defined as objects of governance in various policy-making areas, but rather how their potential future characteristics could render them objects of governance in policy areas where they were either not considered before or were considered in a different manner. This might even lead to the emergence of completely new policy areas. In that regard, it is important that robots be explicitly considered an emerging technology, as will be argued below.

Governance of Emerging Technology and Its Difficulties

The governance of new technology is often based on the assumption that a technology is developed first, after which policy-making initiatives are created to govern its implementation in society so as to regulate certain uses of that technology. Even though many concepts and theoretical frameworks of technological development have argued against this assumption in different ways, it remains a rather stubborn notion. In addition, it can also be connected to a more fundamental problem regarding the character of governance versus the character of technological development. An often-cited and well-defined expression of this problem is the Collingridge dilemma, which still functions as an important reference in fields like responsible research and innovation (RRI) and technology assessment (TA) (Genus and Stirling, 2018). This dilemma was defined by David Collingridge in his 1980 book "The social control of technology," with the book's preface providing a concise and clear definition: "By the time undesirable consequences are discovered, however, the technology is often so much part of the whole economic and social fabric that its control is extremely difficult" (Collingridge, 1982, 11). Particularly in the current moment, which is characterized by technological changes that are changing socioeconomic and political realities in a rapid and profound manner, the dilemma of control is often felt to be particularly prevalent. Examples are multiple, but a prominent one has been the use of big data analytics on social media (e.g. for election campaigns). It is therefore not surprising that calls for a change of approach to technology governance are particularly strong at the moment (Bratton, 2015; OECD, 2017; Schwab, 2017; Winfield and Jirotka, 2018).

The governance of emerging technologies presents an important challenge that has been addressed in different ways in various social science and humanities disciplines. It has been repeatedly noted that the governance of emerging technologies can be seen as quite a specific type of governance (Kuhlmann et al., 2019; Ulnicane et al., 2021). Based on the discussion and analysis of different emerging technologies throughout the years, a useful body of literature has developed discussing the particular status of emerging technologies in policy-making. (Bonnin Roca et al., 2017; Dorbeck-Jung and Bowman, 2017; Kaebnick and Gusmano, 2018). First of all, as previously mentioned, emerging technologies often have the potential to *cause effects on a broad scale* in society (Rotolo et al., 2015). An important issue for the governance of emerging technologies like robotics is that initially relatively small-scale projects can have severe ramifications in the near future, not least because financing schemes in the startup economy render high-risk/high-reward ventures more likely (McNeill, 2016). When it comes to social robots specifically, the main issue concerns their increasing ability to participate in different parts of social life. As demonstrated in the section above, many philosophers have been discussing potential consequences for the organization of social life, and the robot standing debate can very much be seen as a part of this larger discussion. Second and related to the first point, policy-making developments regarding emerging technologies are generally characterized by *widely divergent expectations* concerning the potential futures of those technologies. Apart from general expectations, this also applies very much to sociotechnical imaginaries in policy-making, as has been repeatedly demonstrated (Kearnes et al., 2006; Vesnic-Alujevic et al., 2016; Rieder, 2018). An important reason for this is that emerging technologies are usually surrounded by hype and various buzzwords. In that sense, it is beneficial to apply some vocabulary from STS research, which has a good track record analyzing emerging technologies in relation to public attitudes and governance. A useful term here is "sociotechnical controversy" (Bonneuil et al., 2008). Central to the notion of sociotechnical controversies and their emergence is that they are continuously in the making and are subject to negotiation processes among different stakeholders. Fields like (global) governance studies and STS have extensively analyzed such processes. Finally and related to the first two points, there is often a strong *public interest* in the (potential) development of emerging technologies. This is an issue that is particularly prominent in the case of emerging technologies and their future trajectories, since emerging technologies are often characterized by many different expectations and speculations regarding their future development. Public attitudes towards the sociotechnical controversies around emerging technologies are therefore usually considered to play an important role in the uptake of these technologies. Autonomous technologies like robots in general and social robots more specifically are a particularly prominent issue in this respect. Their (potential) autonomy has been a recurring major theme in many different kinds of media and art for many years, while recent developments in AI technology could indeed bring about a strong leap in the actual autonomy of robotic devices.

Emerging Technologies as Objects of Governance

Above it has become clear that robots, seen as an emerging technology, are to be understood as a challenge in terms of governance. Furthermore, when it comes to issues of governance, it is important to note that emerging technologies suffer from a particularly strong form of fuzziness about their status as objects of governance. This very much applies to emerging robotics (and AI) as well. Central to this problem are challenges regarding the contingencies when it comes to robots as objects of governance. Those contingencies can be understood in two different ways. The first concerns future contingency and is the most straightforward: uncertainty about future technological developments makes technology governance a difficult issue. We do not yet know the future of robots as objects of governance, but want to anticipate it in order to implement governance measures in a timely manner. The second concerns ontological contingencies regarding the object of governance itself. The question here relates to the phenomena that are considered to be part of robots, as well as the different ways those phenomena can be rendered governable. Robots are as such a particularly fuzzy and dynamic phenomenon that is difficult to fully grasp through the different policy-making instruments that are available or could potentially be developed in the future.

In both of these cases of fuzziness, speculative concepts can be instrumental in the constitution of objects of governance by rendering them more explicit. That is to say, by carefully developing arguments on the basis of speculative concepts such as robot standing, we render the (perceived) autonomy and agency of robots into explicit phenomena that define robots in their social context. Relevant here is how speculative concepts can influence the way in which emerging technologies become constituted as future objects of governance. I argue here that it is exactly the speculative element that can help in the further development of anticipatory robotics governance. In this way, the role of forecasting practices as well as policy instruments in general can evolve, especially when it comes to specific technological trends like the emergence of new types of robots. As demonstrated, for instance, by the European Parliament's notion of electronic personality, this type of governance is experiencing continuous evolution as new policy ideas gradually develop.

As already explained above, when it comes to robotics, applied ethics fields like robot ethics have gained influence in policy-making discourse around emerging technologies in recent years. From a governance perspective, this can be seen as a way to anticipate future changes (Brey, 2012). The goal here is thus to develop a better understanding of how technologies like robots become constituted as objects of governance and subsequently elaborate how approaches to future contingencies in the governance of technology are materialized during this process. This will be instrumental for the subsequent discussion section, which further elaborates how robot standing and its speculative content can play a role in the anticipation of autonomous social robots. The analysis of policy-making efforts around unpredictable issues with a high level of controversy and a strong presence of buzzwords has developed considerably in recent decades, especially in STS research (Fortun, 2001; Hilgartner, 2009). In that regard, it is useful to elaborate on robot governance by drawing upon literature from this field and other policy research around the notion of "objects of governance". Other terms that are often used in this context are "governance object" or "object of government" (Lezaun, 2006). When used as a concept for analysis, an important assumption is that governance arrangements around objects of governance can be traced back to contested representations in earlier phases of their emergence as objects of governance (Allan, 2017). The underlying idea is that objects of governance are hybrid, co-produced entities that emerge from complex interactions between expert knowledge, political interventions and mundane practices (Allan, 2018). In other fields of research, it has already been demonstrated how epistemic communities play a central role in the development of new and altered policy ideas (Swinkels, 2020). Examples of such research are: the climate as an object of (global) governance (Bulkeley, 2005; Allan, 2017), urban warming as an object of (local) governance (Boezeman and Kooij, 2015), or creative thinking as an object of governance and geopolitical concern in the United States military context during the Cold War (Van Eekelen, 2017). As such studies show, anything can become a governance object as long as it becomes distinguishable and is rendered governable. Bentley Allan provides a comprehensive description of governance objects when he defines them as "concatenations of knowledges, artifacts, physical phenomena, and practices that have been yoked together and constituted as an entity distinct from other objects, events, and actors" (2018, 13). By applying his perspective, networks can be understood in a way that allows for high levels of complexity and contingency. Furthermore, the process of such networks' emergence and stabilization is of great interest to policy researchers in the sense that new networks of cooperation are developed to link elements that were previously disconnected (Jessop, 2011). Therefore, a crucial part of the theory behind the analysis of objects of governance is the notion that how objects of governance become defined as such is dependent on negotiation processes underlying sociotechnical controversies. A major quality of this approach is its capacity to explain how and why a specific version of an object of governance emerges. Such an analysis can be very useful because it helps provide new insights into the dynamic processes and (path-dependent) characteristics of technoscientific governance. Finally, the fact that this approach is very much open to novel, emergent understandings of the object of governance at hand can be quite useful. Instead of understanding robotic technologies as something pre-defined, the goal is to look at the way in which it is exactly the above-mentioned processes of interaction that are responsible for their constitution as an object of governance. The approach of analyzing new phenomena as objects of governance (or comparable concepts) is useful for social scientists because of its possibilities for applying a critical perspective: by developing an understanding of underlying governance processes, it becomes feasible to criticize their assumptions.

Nevertheless, there is a difference between the approach to objects of governance described above and the objective of this paper. The different studies mentioned above focus on (recent) pasts: they trace, often through qualitative empirical social research, how something emerges as an object of governance. This paper is neither focused on tracing the (recent) past of robotics governance, nor does it aim to systematically present the outcomes of empirical social research. Rather, it seeks to develop an understanding of robot standing as a speculative concept while conceptualizing its contribution to the process of robots becoming objects of future governance. In other words, the object of governance concept is used to exploratively establish the role of speculative concepts like robot standing in the governance of (social) robots, rather than descriptively criticizing existent and past robotics governance. As such, the paper focuses more strongly on the mission of philosophy of technology rather than the social sciences when it comes to these matters. In a more general sense, the argument here is that the systematic and robust application of speculative concepts can aid the process of constituting better, more profound objects of future governance that aid the process of implementing robots into our society in a sustainable manner. As previously stated, complex objects of governance by default go through different processes of negotiation along the lines of epistemic disagreements. Therefore, on a governance level, if philosophers (of technology) are provided with the possibility to engage with the development of policy ideas and demonstrate their insights, they can be participants in the negotiation processes behind sociotechnical controversies, with their concepts serving as their currency. In light of this, the section below will explain why and how robot standing can be seen as such a concept by framing the issue of robot standing as an important rhetorical and analytical device in the process of constituting robots as objects of governance.

ROBOT STANDING AND THE GOVERNANCE OF SOCIAL ROBOTS

The preceding sections have explained how robot standing can be understood as a speculative concept that can aid the process of negotiating how (social) robots are to be constituted as objects of governance. The subsections below explore different uses of the robot standing concept in more detail. They describe the ramifications of applying the concept of robot moral standing in discussions on the futures of robots. In doing so, my aim is to develop some concrete insights and proposals of how a speculative concept like robot standing can be of help in the deliberative processes behind the development of new policy ideas. This should help to determine how some futures might be prevented so that other futures can be realized (Bratton, 2021). Three different points are distinguished: the understanding of social robots, analysis of robots' societal impact, and the exploration of (social) robots' sociotechnical potential.

Facilitating New Understandings of Social Robots

Part and parcel of the analysis of the process in which objects of governance become constituted is the idea that specific policy ideas and the concepts related to them, are important for enabling governance in a volatile, high-stakes context (Schaper-Rinkel, 2013). However, from a governance standpoint, it is certainly impossible to track down every small-scale but potentially large-impact instance of technological development from the start and understand its consequences. What can be done is to develop different guiding concepts and narratives that are sufficiently broad while avoiding deterministic views of technological development. In such a context, the speculative endeavour towards concepts of moral standing can be described as attempts to provide more sophisticated understandings of social robot morality as such. Because of its disciplinary focus on the development of concepts and conceptual schemes, philosophy of technology plays an important role in developing those understandings. In recent decades, philosophy of technology and related fields have seen quite a transformation, which is often referred to as an "empirical turn" (Brey, 2010). Now that this turn has become quite established, the question is in which ways philosophy of technology should aim to influence policy ideas and improve the concepts that can be used in the negotiation processes behind the constitution of (social) robots as objects of governance. Since philosophers (of technology) have a great track record concerning the moral and mental standing of humans and other beings, it is desirable that they continue such activities. Whereas artificial concerns with no ground in engineering reality should probably be avoided, it is also important to actively learn what kind of speculative concepts have the ability to support the development of more sophisticated and profound understandings of robots as objects of governance. The question is therefore not whether we should have a concept of robot standing, but rather, what kind of concepts of robot standing we want to explore and which ones should better be set aside. Naturally, interdisciplinary and transdisciplinary interactions are crucial here in order to continuously discuss the (ir)relevance of specific concepts, tweak their definitions, and explore their potential ramifications.

When philosophers explore new ontologies and identify lacunae within existing ontologies, the goal is to create new understandings of the demarcation and definition of the meaning of robots in specific contexts. In this way, they can demonstrate the ways in which robots can disturb existing ontologies. Crucial here is that associated concepts can be applied in different contexts. Philosophical elaborations on such changes can in that way become relevant to many other academic disciplines, such as law, HRI, and critical governance studies. For example, in his abovementioned work on climate as an object of governance, Bentley Allan describes how the notion of the climate in governance shifted from a bioecological to a geophysical understanding, because "US state agencies drove billions of dollars into the institutions of knowledge production, altering their priorities, trajectories, and products" (Allan, 2017, 157). In the same way, social robotics is currently becoming defined via specific priorities, trajectories and products. In this regard, automation should not only be understood as the outcome of engineering inventions. It is also something that must be discovered in its context of development. Philosophers can help shape the debate around such phenomena so that they can be understood in new and better ways. As this paper has argued, the use of well-developed speculative concepts is instrumental in such forms of engagement. The role of the philosopher is thus not necessarily to speculate continuously. Instead, it is to engage with speculative concepts and apply philosophical rigour to their potential ramifications. Even though fully autonomous social robots are still far from being realized, it is important to engage with their technological potential in a rigorous manner so as to facilitate the new understandings of social robots and their roles within the social contexts in which they will play unprecedented roles.

Enabling Critical Long-Term Analysis of Robots' Societal Impact

The main use of the object of governance notion as an analytical tool is to capture how policy-making takes shape. This lens demonstrates how investing in speculative concepts can be instrumental for the constitution of new objects of governance. A major advantage of this is that the use of such concepts will make it possible to trace different societies' views and narratives concerning those concepts over a longer period of time. Debates on robot standing, as a key example, will most certainly change considerably over time. Having this concept available renders it a possibility for social scientists to analyze the discourses and narratives around robot standing. Directly related to this, it is important to analyze what can be done to make the moral and sociopolitical assumptions behind robots as transparent as possible. From a governance perspective, it is therefore useful to look at robots as artificial social agents and establish in which ways the artificial sociality of robots can become defined. In this way, the analysis and decision-making processes concerning the impact of robots can become more pluralistic. For our analysis of the impact of robotics and other emerging technologies, we are currently still too often depending on analytical tools that have been criticized for years for their lack of nuance. For example, the effects of robotic technology on society and the economy in order to facilitate governmental decision-making are still mostly analyzed via quantitative, mostly macroeconomic indicators that measure the effect of robots and automation on a country's GDP, its employment rate and so on (National Academies of Sciences, Engineering, and Medicine, 2017). Future-oriented concepts make it possible to analyze the effect of (social) robots in the long term in different (qualitative) ways, since changes in meaning can be traced with discursive methods, as demonstrated by the examples of research on objects of governance. The advent of social robots in an increasingly complex society full of contradictory regimes of information makes it important to improve this type of analysis.

Therefore, even though quantitative indicators will remain important, and rightly so, it is useful to aim for speculative concepts that are likely to remain relevant for a longer period of time and are based on both social and technical contexts. The choice of such concepts is not easy and will certainly include concepts and ideas that fade away later, as they will turn out to be unfit for how technological development actually comes about. Therefore, which concepts qualify as useful in this context and which do not will always be a point of discussion. This paper argues that robot standing can be seen as a useful concept because it engages with the potentialities of robotics while being clearly linked to both cultural fascinations and ethical and legal systems. Furthermore, qualitative and quantitative indicators can be used together to improve the analysis of how automated social robots can be implemented in social life. Such concepts are critical for engaging with the future, particularly for research in the social sciences and humanities. Once such concepts become established, it not only becomes possible to have informed discussions about potential characteristics of robots, but also to trace how such concepts develop in the long term. This allows social research to monitor and map the sociocultural notions regarding technologies like social robotics in a more credible and structural manner. In order to do that, we need concepts that can help to analyze a specific sociotechnical controversy in a rigorous manner (Marres, 2015). In this way, we will hopefully be able to improve our understanding of long-term dynamics in large-scale sociotechnical systems.

Exploring Social Robotics' Sociotechnical Potential

Finally, a concept like robot standing also allows for explorative imagination of the future as a way of motivating new, emancipatory social ontologies (Lewis et al., 2018). In this context, speculative explorations of robot moral standing can be used to analyze moral and legal adaptations to potential future characteristics of the social fabric. Generally speaking, the type of imaginative thinking that serves as a foundation for ideas for sociopolitical change has historically been an important element of ideas and concepts in the humanities and social sciences. For instance, in recent decades, posthumanist thinking has been an important field that has mobilized the technoscientific imagination in order to argue for new, more equal sociopolitical realities. Crucial to such contemplations of the posthuman being as a political subject are the fact that they do not need to reach the status of material reality. Important examples of such sociopolitical entities include the cyborg in Donna Haraway's *A Cyborg Manifesto*, which was already mentioned in above. Another more recent recent example is Aaron Bastani's *Fully Automated Luxury Communism* (Bastani, 2019). In what he explicitly calls a "manifesto", Bastani calls for full automation and common ownership of that which is being automated. In certain ways, the discussion on robot rights and robot standing has already contributed to comparable issues. Two different examples can help to illustrate this, the first being the robot Sofia, which received citizen rights in Saudi Arabia, which in turn sparked several discussions on how a robot apparently has more rights in Saudi Arabia than other minorities. Another example closer to philosophy of technology comes from Kathleen Richardson, who presents a firm argument by claiming that many of the discussions concerning changing human-machine boundaries and associated calls for robot rights and standing merely appear to be progressive, while are in fact based on the old but persistent (Arestotelian) notions of humans as property. In her argument, granting rights to robots is synonymous to granting rights to slaves, which then serves as a way to ignore modern forms of human slavery in general (Richardson, 2015, 2016). Even though several of the arguments in this paper at least partly contradict Richardson's ideas, it is important to appreciate the clarity and firmness of her arguments on anti-essentialism and its relation to the rejection of ontological differences between humans and machines. In this way, the powerful imagery of the social robot can lead to important discussions on human sociality.

Hence, I argue here that the social robot can be used as a point of sociopolitical reflection and imagination. This is certainly not a new argument. For instance, as Scott Selisker nicely describes in his study of the human automaton in American politics, such imagery of the automaton became a common trope in portrayals of totalitarian governance while also figuring as an important element in progressive accounts of future societies (Selisker, 2016). In the same way, with the help of imagery of potential technological developments, autonomous social robots can already be imagined as sociopolitical agents, even though they might never become actual reality. Looking at social robots as objects of future governance in this way means that current social ontologies are continuously scrutinized (Sætra, 2021). As our legal and ethical systems and values need to be critically reviewed in this process, powerful concepts like moral standing can be used as rhetorical devices that enable specific understandings of human vs robot moral standing in the negotiation space for values surrounding (social) robots. Rather than resistance against robots as such, ethical and judiciary concepts can be developed as robust and innovative instruments for debates that aim to create more equal futures with robots. Those utopian social ontologies can then be applied to criticize actual governance, particularly in light of ironic and subversive elements in their argumentation. Imagery of the posthuman other is often a simultaneously fascinating as well as daunting prospect. Nevertheless, from a governance perspective, it might be tempting to equate such efforts to the hubris that surrounds emerging technologies in general. In fact, in addition to the discussions on the potential effects of robotics on very fundamental habits, they stimulate and obligate important discussions on crucial concepts lying under the surface of society.

CONCLUSION

Philosophy of technology has already made considerable efforts towards increasing involvement in the development of policy ideas. This paper has aimed to provide several arguments about how the speculative element of such efforts can be beneficial for the process of constituting social robots as objects of governance in an intelligent and informed manner. This paper has argued that the development of a concept like robot standing should be understood as an effort to develop concepts that are speculative but rigorous. Both are required to achieve this goal, which will also necessitate efforts to develop such concepts further while testing their usefulness outside of philosophy. With respect to the new normal of emerging technology, part of the solution can be found in the development of new idioms and imaginaries that can help to understand new technology and how its different futures (e.g. technological, social, political, economic) are incoherent with each other. Thus, it is important that speculative futures concerning emerging technologies be taken seriously and engaged with. Rather than understanding the technological future as a fantasmatic projection, the idea is to engage critically with it and its narratives. This also means that instead of disapproving of the future-grasping, speculative character of technological visions, there is a need to invest rather more than less into speculative concepts like robot standing. It is through the thorough analysis of these concepts that philosophy of technology can actively participate in the prescriptive engagement with technology futures.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

Open access funding provided by TU Wien (TUW).

ACKNOWLEDGMENTS

This work was supported by the TrustRobots Doctoral College, TU Wien.

REFERENCES

AIHLEG (2019). Ethics Guidelines for Trustworthy AI. Available at: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

Alač, M. (2016). Social Robots: Things or Agents? *AI Soc.* 31, 519–535. doi:10.1007/s00146-015-0631-6

Allan, B. B. (2018). From Subjects to Objects: Knowledge in International Relations Theory. *Eur. J. Int. Relations* 24, 841–864. doi:10.1177/1354066117741529

Allan, B. B. (2017). Producing the Climate: States, Scientists, and the Constitution of Global Governance Objects. *Int. Org.* 71, 131–162. doi:10.1017/S0020818316000321

Bastani, A. (2019). *Fully Automated Luxury Communism*. London; New York: Verso Books.

Bertolini, A., and Aiello, G. (2018). Robot Companions: A Legal and Ethical Analysis. *Inf. Soc.* 34, 130–140. doi:10.1080/01972243.2018.1444249

Birhane, A., and van Dijk, J. (2020). Robot Rights? Let's Talk about Human Welfare Instead In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (New York: ACM)*, 207–213. doi:10.1145/3375627.3375855

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of Robotics: Regulating Robots in the Real World. *Connect. Sci.* 29, 124–129. doi:10.1080/09540091.2016.1271400

Boezeman, D., and Kooij, H. J. (2015). Heated Debates: the Transformation of Urban Warming into an Object of Governance in the Netherlands. in *Evolutionary Governance Theory*. New York: Springer, 185–203. doi:10.1007/978-3-319-12274-8_13

Bonneuil, C., Joly, P.-B., and Marris, C. (2008). Disentrenching Experiment. *Sci. Technol. Hum. Values* 33, 201–229. doi:10.1177/0162243907311263

Bonnín Roca, J., Vaishnav, P., Morgan, M. G., Mendonça, J., and Fuchs, E. (2017). When Risks Cannot Be Seen: Regulating Uncertainty in Emerging Technologies. *Res. Pol.* 46, 1215–1233. doi:10.1016/j.respol.2017.05.010

Bösl, D. B. O., and Bode, M. (2018). Roboethics and Robotic Governance - A Literature Review and Research Agenda. in *ROBOT 2017: Third Iberian Robotics Conference*. Editors A. Ollero, A. Sanfeliu, L. Montano, N. Lau, and C. Cardeira (Cham: Springer International Publishing), 140–146. doi:10.1007/978-3-319-70833-1_12

Bratton, B. H. (2015). *The Stack: On Software and Sovereignty*. Cambridge, Massachusetts: MIT Press.

Bratton, B. (2021). "New World Order": For Planetary Governance. *Strelka Mag.* Available at: https://strelkamag.com/en/article/new-world-order-for-planetary-governance (Accessed June 6, 2021).

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *Nanoethics* 6, 1–13. doi:10.1007/s11569-012-0141-7

Brey, P. (2010). Philosophy of Technology after the Empirical Turn. *Techné: Res. Philos. Tech.* 14, 36–48. doi:10.5840/techne20101416

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: the Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25, 273–291. doi:10.1007/s10506-017-9214-9

Bryson, J. J. (2010). Robots should be slaves, in *Close engagements with artificial companions: key social, psychological, ethical and design issues*, ed. Y. Wilks (Amsterdam: John Benjamins Publishing Company), 63–74. Available at: https://researchportal.bath.ac.uk/en/publications/robots-should-be-slaves (Accessed June 24, 2019).

Bulkeley, H. (2005). Reconfiguring Environmental Governance: Towards a Politics of Scales and Networks. *Polit. Geogr.* 24, 875–902. doi:10.1016/j.polgeo.2005.07.002

Chan, K. M. A. (2011). Ethical Extensionism under Uncertainty of Sentience: Duties to Non-human Organisms without Drawing a Line. *Environ. Values* 20, 323–346. doi:10.3197/096327111X13077055165983

Coeckelbergh, M. (2010). Robot Rights? towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12, 209–221. doi:10.1007/s10676-010-9235-5

Coeckelbergh, M. (2014). The Moral Standing of Machines: Towards a Relational and Non-cartesian Moral Hermeneutics. *Philos. Technol.* 27, 61–77. doi:10.1007/s13347-013-0133-8

Coeckelbergh, M. (2015). The Tragedy of the Master: Automation, Vulnerability, and Distance. *Ethics Inf. Technol.* 17, 219–229. doi:10.1007/s10676-015-9377-6

Collingridge, D. (1982). The Social Control of Technology. *Repr*. London: Pinter Publishers.

Cressman, D. (2020). Contingency and Potential: Reconsidering a Dialectical Philosophy of Technology. *Techné: Res. Philos. Tech.* 24, 1–20. doi:10.5840/techne202027114

Dafoe, A. (2015). On Technological Determinism. *Sci. Technol. Hum. Values* 40, 1047–1076. doi:10.1177/0162243915579283

Danaher, J. (2019). The Rise of the Robots and the Crisis of Moral Patiency. *AI Soc.* 34, 129–136. doi:10.1007/s00146-017-0773-9

Dautenhahn, K. (2007). Socially Intelligent Robots: Dimensions of Human-Robot Interaction. *Phil. Trans. R. Soc. B* 362, 679–704. doi:10.1098/rstb.2006.2004

Dorbeck-Jung, B., and Bowman, D. M. (2017). Regulatory Governance Approaches for Emerging Technologies. In *Embedding New Technologies into Society*. Boca Raton: Jenny Stanford Publishing, 35–59. doi:10.1201/9781315379593-3

EP (2017). Civil Law Rules on Robotics - European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). Available at: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html (Accessed January 5, 2020).

Floridi, L., and Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14, 31. doi:10.1023/b:mind.0000035461.63578.9d

Fortun, M. (2001). Mediated Speculations in the Genomics Futures Markets. *New Genet. Soc.* 20, 139–156. doi:10.1080/14636770124557

Fosch-Villaronga, E., Lutz, C., and Tamò-Larrieux, A. (2020). Gathering Expert Opinions for Social Robots' Ethical, Legal, and Societal Concerns: Findings from Four International Workshops. *Int. J. Soc. Robotics* 12, 441–458. doi:10.1007/s12369-019-00605-z

Gellers, J. C. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. 1st ed. Routledge. doi:10.4324/9780429288159

Genus, A., and Stirling, A. (2018). Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation. *Res. Pol.* 47, 61–69. doi:10.1016/j.respol.2017.09.012

Grunwald, A. (2010). From Speculative Nanoethics to Explorative Philosophy of Nanotechnology. *Nanoethics* 4, 91–101. doi:10.1007/s11569-010-0088-5

Gunkel, D. J. (2018a). *Robot Rights*. Cambridge, Massachusetts: MIT Press.

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, Mass: MIT Press.

Gunkel, D. J. (2018b). The Other Question: Can and Should Robots Have Rights? *Ethics Inf. Technol.* 20, 87–99. doi:10.1007/s10676-017-9442-4

Guston, D. H. (2014). Understanding 'anticipatory Governance. *Soc. Stud. Sci.* 44, 218–242. doi:10.1177/0306312713508669

Haraway, D. (1991). A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In *Simians, Cyborgs and Women: The Reinvention of Nature*. New York: Routledge, 149–181.

Harman, G. (2016). *Immaterialism: Objects and Social Theory*. Cambridge: Polity Press.

Heffernan, T. (2019). "Fiction Meets Science: Ex Machina, Artificial Intelligence, and the Robotics Industry," in *Cyborg Futures*. Editor T. Heffernan (Berlin: Springer), 127–140. doi:10.1007/978-3-030-21836-2_7

Heilbroner, R. (1994). Technological determinism revisited, in *Does Technology Drive History? The Dilemma of Technological Determinism, eds. M. R. Smith and L. Marx (Cambridge, MA: MIT Press)*, 67–78.

Hilgartner, S. (2009). Intellectual Property and the Politics of Emerging Technology: Inventors, Citizens, and Powers to Shape the Future. *Chicago-Kent L. Rev.* 84, 197-224.

Jessop, B. (2011). Metagovernance. In *The SAGE Handbook of Governance*. Los Angeles, CA: SAGE Publications, 106–123.

Kaebnick, G. E., and Gusmano, M. K. (2018). Making Policies about Emerging Technologies. *Hastings Cent. Rep.* 48, S2–S11. doi:10.1002/hast.816

Kearnes, M., Grove-White, R., Macnaghten, P., Wilsdon, J., and Wynne, B. (2006). From Bio to Nano: Learning Lessons from the UK Agricultural Biotechnology Controversy. *Sci. as Cult.* 15, 291–307. doi:10.1080/09505430601022619

Kim, M.-S., and Kim, E.-J. (2013). Humanoid Robots as "The Cultural Other": Are We Able to Love Our Creations? *AI Soc.* 28, 309–318. doi:10.1007/s00146-012-0397-z

Koops, B.-J., Di Carlo, A., Nocco, L., Casamassima, V., and Stradella, E. (2013). Robotic Technologies and Fundamental Rights. *Int. J. Technoethics* 4, 15–35. doi:10.4018/jte.2013070102

Kuhlmann, S., Stegmaier, P., and Konrad, K. (2019). The Tentative Governance of Emerging Science and Technology-A Conceptual Introduction. *Res. Pol.* 48, 1091–1097. doi:10.1016/j.respol.2019.01.006

Latour, B. (2014). Agency at the Time of the Anthropocene. *New Literary Hist.* 45, 1–18. doi:10.1353/nlh.2014.0003

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford ; New York: Oxford University Press.

Law, J. (2008). On Sociology and STS. *Sociological Rev.* 56, 623–649. doi:10.1111/j.1467-954X.2008.00808.x

Leenes, R., Palmerini, E., Koops, B.-J., Bertolini, A., Salvini, P., and Lucivero, F. (2017). Regulatory Challenges of Robotics: Some Guidelines for Addressing Legal and Ethical Issues. *L. Innovation Tech.* 9, 1–44. doi:10.1080/17579961.2017.1304921

Lewis, J. E., Arista, N., Pechawis, A., and Kite, S. (2018). Making Kin with the Machines. *J. Des. Sci.* 3.5. doi:10.21428/bfafd97b

Lezaun, J. (2006). Creating a New Object of Government. *Soc. Stud. Sci.* 36, 499–531. doi:10.1177/0306312706059461

Lindstrøm, T. C. (2015). Agency 'in itself'. A Discussion of Inanimate, Animal and Human agency. *Arch. Dial.* 22, 207–238. doi:10.1017/S1380203815000264

Marres, N. (2015). Why Map Issues? on Controversy Analysis as a Digital Method. *Sci. Technol. Hum. Values* 40, 655–686. doi:10.1177/0162243915574602

Maulsby, D., Greenberg, S., and Mander, R. (1993). Prototyping an Intelligent Agent through Wizard of Oz. In Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems CHI '93. New York, NY, USA: Association for Computing Machinery, 277–284. doi:10.1145/169059.169215

McNeill, D. (2016). Governing a City of Unicorns: Technology Capital and the Urban Politics of San Francisco. *Urban Geogr.* 37, 494–513. doi:10.1080/02723638.2016.1139868

Mindell, D. A. (2015). Our Robots, Ourselves: Robotics And the Myths of Autonomy. *Viking Adult*.

National Academies of Sciences, Engineering, and Medicine (2017). *Information Technology and the U.S. Workforce: Where Are We and where Do We Go from Here* Washington, DC: The National Academies Press. doi:10.17226/24649

Nevejans, N. (2018). Open Letter to the European Commission: Artificial Intelligence and Robotics. Available at: http://www.robotics-openletter.eu.

Nordmann, A. (2007). If and Then: A Critique of Speculative NanoEthics. *Nanoethics* 1, 31–46. doi:10.1007/s11569-007-0007-6

Nordmann, A. (2014). Responsible Innovation, the Art and Craft of Anticipation. *J. Responsible Innovation* 1, 87–98. doi:10.1080/23299460.2014.882064

Nordmann, A., and Rip, A. (2009). Mind the gap Revisited. *Nat. Nanotech* 4, 273–274. doi:10.1038/nnano.2009.26

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci. *Sci. Eng. Ethics* 24, 1201–1219. doi:10.1007/s11948-017-9943-x

OECD (2017). *Trust and Public Policy: How Better Governance Can Help Rebuild Public Trust*. Paris: OECD Publishing. doi:10.1787/9789264268920-en

Pagallo, U. (2013). *The Laws of Robots*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-6564-1

Pasquale, F. (2020). *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Harvard University Press. doi:10.4159/9780674250062

Pickering, A. (2017). The Ontological Turn: Taking Different Worlds Seriously. *Soc. Anal.* 61. doi:10.3167/sa.2017.610209

Richardson, K. (2015). *An Anthropology of Robots and AI: Annihilation Anxiety and Machines*. Routledge. doi:10.4324/9781315736426

Richardson, K. (2016). Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines. *IEEE Technol. Soc. Mag.* 35, 46–53. doi:10.1109/MTS.2016.2554421

Rieder, G. (2018). *"Tracing Big Data Imaginaries through Public Policy: The Case of the European Commission," in the Politics of Big Data - Big Data, Big Brother* Routledge, 89–109.

Riek, L. (2012). Wizard of Oz Studies in HRI: a Systematic Review and New Reporting Guidelines. *Jhri* 1, 119–136. doi:10.5898/JHRI.1.1.Riek

Rossi, C., Russo, F., and Russo, F. (Editors) (2009). "Automata (Towards Automation and Robots)," in Ancient Engineers& Inventions History of Mechanism and Machine Science (Dordrecht: Springer Netherlands), 269–301. doi:10.1007/978-90-481-2253-0_15

Rotolo, D., Hicks, D., and Martin, B. R. (2015). What Is an Emerging Technology? *Res. Pol.* 44, 1827–1843. doi:10.1016/j.respol.2015.06.006

Sætra, H. S. (2021). Robotomorphy. *AI Ethics*. doi:10.1007/s43681-021-00092-x

Schaper-Rinkel, P. (2013). The Role of Future-Oriented Technology Analysis in the Governance of Emerging Technologies: The Example of Nanotechnology. *Technol. Forecast. Soc. Change* 80, 444–452. doi:10.1016/j.techfore.2012.10.007

Schwab, K. (2017). *The Fourth Industrial Revolution*. New York: Crown Publishing Group.

Selin, C. (2014). On Not Forgetting Futures. *J. Responsible Innovation* 1, 103–108. doi:10.1080/23299460.2014.884378

Selisker, S. (2016). *Human Programming: Brainwashing, Automatons, and American Unfreedom*. Minneapolis: University of Minnesota Press.

Suchman, L. (2019). "Demystifying the Intelligent Machine," in Cyborg futures: cross-disciplinary perspectives on artificial intelligence and robotics. Editor T. Heffernan (Berlin: Springer), 35–61. doi:10.1007/978-3-030-21836-2_3

Suchman, L. (2006). *Human–Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press.

Swinkels, M. (2020). How Ideas Matter in Public Policy: a Review of Concepts, Mechanisms, and Methods. *irpp* 2, 281–316. doi:10.4000/irpp.1343

Taeihagh, A. (2021). Governance of Artificial Intelligence. *Pol. Soc.* 40, 137–157. doi:10.1080/14494035.2021.1928377

Turner, J. (2019). *Robot Rules: Regulating Artificial Intelligence*. Basingstoke: Palgrave Macmillan.

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., and Wanjiku, W.-G. (2021). Framing Governance for a Contested Emerging Technology:insights from AI Policy. *Pol. Soc.* 40, 158–177. doi:10.1080/14494035.2020.1855800

Van de Poel, I. (2016). An Ethical Framework for Evaluating Experimental Technology. *Sci. Eng. Ethics* 22, 667–686. doi:10.1007/s11948-015-9724-3

Van Eekelen, B. F. (2017). Creative Intelligence and the Cold War. *Conflict Soc.* 3, 92–107. doi:10.3167/arcs.2017.030108

Vesnic-Alujevic, L., Breitegger, M., and Pereira, Â. G. (2016). What Smart Grids Tell about Innovation Narratives in the European Union: Hopes, Imaginaries and Policy. *Energ. Res. Soc. Sci.* 12, 16–26. doi:10.1016/j.erss.2015.11.011

Wallach, W. (2015). *A Dangerous Master: How to Keep Technology from Slipping beyond Our Control*. Hachette UK.

Wallach, W., and Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford ; New York: Oxford University Press.

Winfield, A. F. T., and Jirotka, M. (2018). Ethical Governance Is Essential to Building Trust in Robotics and Artificial Intelligence Systems. *Phil. Trans. R. Soc. A.* 376, 1–13. doi:10.1098/rsta.2018.0085

Wyatt, S. (2008). "Technological Determinism Is Dead; Long Live Technological Determinism," in *The Handbook of Science & Technology Studies*. Editors E. J. Hackett, O. Amsterdamska, M. Lynch, and J. Wajcman (Cambridge, MA: MIT Press), 165–181.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

HYPOTHESIS AND THEORY
published: 23 December 2021
doi: 10.3389/frobt.2021.789327

Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence

Sergio M. C. Avila Negri *

Department of Private Law, Federal University of Juiz de Fora, Juiz de Fora, Brazil

This paper seeks to investigate the proposal to create a legal (electronic) personhood for robots with artificial intelligence based on the European Parliament resolution with recommendations on Civil Law and Robotics. To this end, we highlight the various risks and problems present in this type of initiative, especially in view of the current trend of expanding legal subjectivity in various jurisdictions. In addition to an anthropomorphic rhetoric, we can observe the prevalence of a pragmatic line that seeks to be guided, mainly, by the model of corporations, without taking into account, however, problems present in the process of embodiment of companies and the particular function of the term legal person in the grammar of Law.

Keywords: artificial intelligence, legal personhood, robotics, electronic personhood, legal person

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Jaana Parviainen,
Tampere University, Finland
Andreas Theodorou,
Umeå University, Sweden

****Correspondence:***
Sergio M. C. Avila Negri
sergio.negri@ufjf.br

Specialty section:
This article was submitted to
Ethics in Robotics and Artificial Intelligence,
a section of the journal
Frontiers in Robotics and AI

Received: *04 October 2021*
Accepted: *02 December 2021*
Published: *23 December 2021*

Citation:
Avila Negri SMC (2021) Robot as Legal Person: Electronic Personhood in Robotics and Artificial Intelligence.
Front. Robot. AI 8:789327.
doi: 10.3389/frobt.2021.789327

INTRODUCTION

In his essay *The Sphere of Pascal*, the writer Jorge Luis Borges reports that the Greek philosopher Xenophanes, master of Parmenides, was tired of the Homeric verses that dressed the Gods as human beings. In opposition to anthropomorphic traits, he proposed to the Greeks one God, who was in fact an eternal sphere. History followed its course and the exaggeratedly human gods were relegated to poetic fictions.

The anthropomorphic metaphor is not restricted to mythical or religious imagery. Sophia, a humanoid robot with Artificial Intelligence (AI), developed by the Hanson Robotics company, received citizenship from Saudi Arabia in 2017. Although several interviewers were impressed with the sophistication of its responses, the robot follows a simple algorithm and most of its statements are credited to a previously prepared text (Parviainen and Coeckelbergh, 2020).

As in Borges' essay, Robotics can also be thought of without any anthropomorphic resource, with other metaphors, as a sphere. Roomba is a flat, round domestic robot. Even though it does not have social skills like Sophia, the fact that this robotic vacuum cleaner moves on its own, following a simple algorithm, causes some people to give it a name, talk to it and feel bad when the appliance gets stuck under the sofa (Darling, 2016).

If, for a long time, the idea that robots and human beings should be separated was in force, an opposite trend has been accentuated, especially in the last decade: human beings can and should share the same environment as robotic artefacts. As escorts of the elderly—and even children with autism—surgical apparatus, deliverers or security guards, robots have already begun to enter people's homes and lives.

Because of the lack of ontological and legal definition about this emerging technology, the Law is forced to resort to old figures, already-known metaphors, which help us to approach with a certain familiarity what is new and unknown. In 2017, the European Parliament put forward a resolution with guidelines on Robotics, with a proposal to create an electronic personhood for "intelligent" robotic artefacts (European Union, 2017).

In the verbalized legal world, the term “legal person” refers to an autonomous centre of legal relations. The ascription of legal personhood is based on the assumptions that all legal relations take place among natural person and artificial legal person, such as corporations. Following that, the term natural person refers to a human being. By contrast, the term “legal person” or “legal entity” will be often used in this paper when referring to the artificial legal person.

According to Gurnkel (2018a, 2018b), it is important to separate certain questions that are confused in the debates about the legal personhood of robots. First, there is a relevant difference between the two verbs that comprise the question: “can” and “should” AI be persons (Gurnkel, 2018a.) On the other hand, there is another relevant difference, between natural person and legal person. Following that, if legal personhood is already dissociated from the human substrate, there would be no way to deny that AI can be a legal person. But, just because it is possible, that does not mean it should be a good idea.

Just as it is important to separate the idea of moral personhood from the concept of legal personality, it should also be noted that the moral community is not limited to the figure of moral agents, currently reaching the figure of moral patients, who are affected by the actions of (“rational”) agents. This means that rights must not be confused with moral personhood. Likewise, courts might recognize certain legal rights without this implying the recognition of a moral personhood or a general legal personality.

In dialogue with these important elements of the debate on the legal personality of robots, I would like to highlight in this work a distinct aspect, very sensitive to the practice of Law: the legal entity presents itself as a decision structure, which allows the identification of problems and normative solutions used in previous cases. In this sense, it is important to understand the heuristic function of the term legal entity, that is, a mental shortcut that allows, with simplified information, quick judgments.

In the debate on electronic personhood, it is commonly observed that the legal person is presented as if there were no problems in the process of attributing legal personhood to corporations and companies. The analogy with the legal person requires, however, an understanding of the function of this term in legal grammar. This paper seeks to investigate the proposal to create a legal (electronic) personhood for robots based on the European Parliament resolution with recommendations on Civil Law and Robotics. To this end, we highlight the various risks and problems present in this type of initiative, especially in view of the current trend of expanding legal subjectivity in various jurisdictions. In addition to an anthropomorphic rhetoric, we can observe the prevalence of a pragmatic line that seeks to be guided, mainly, by the model of corporations, without taking into account, however, problems present in the process of embodiment of companies and the particular function of the term legal person in the grammar of Law.

PRIVATE LAW AND ROBOTICS

The architecture of digital platforms is capable, in certain cases, of influencing society more directly and efficiently than Law itself. In the growing scenario of technical regulation, it is important to note that programmers and engineers may have difficulty translating ethical and fundamental values into demands that decisively affect people’s lives. In this sense, Langdon Winner (1985), as Leenes (2011) recalls, was already working with the political dimension of artefacts and cited, for example, the absurd, structurally elitist urban constructions of Robert Moses in New York, which were designed to physically impede the passage of public transport to noble areas of the city, since it was predominantly used by the black population.

With the emergence of cyberspace, Information and Communication Technologies (ICTs) have come to be understood as instruments capable of conditioning behaviours. The relationship between Law and the normative effects of technology has been consolidated as a field of study. Lessig (2006) presents the “code”—in his words, the hardware and software that make up cyberspace—as a new form of regulation, since it defines the terms in which interactions in cyberspace take place. Thus, as the code changes, so does the character of cyberspace. Technology always incorporates certain rules, which allow a certain behaviour and inhibit another. Therefore, the rules in cyberspace are increasingly shaped by technology rather than by Law.

Robotics cannot be seen as a novelty. In industry, with emphasis on automobile manufacturing, Robotics represents a technique already incorporated into production, mainly in relation to the performance of routine tasks. As Pagallo (2018) points out, more than 50 years ago “robots have already materialised as a reprogrammable machine, operating semi or fully automatically in manufacturing operations and other industrial tasks” (Pagallo, 2018). Although Robotics should not be confused with AI, it is undeniable that today these fields are more and more closely intertwined, mainly due to the improvement of probabilistic methods, the increasing availability of huge amounts of data and the increase in computational power. One cannot forget, either, the more recent transformation of places and spaces into environments more receptive to information technology, as occurs with the imagery of intelligent cities.

The European Parliament Resolution of February 16, 2017 established that a robot shall be considered intelligent when it has the following characteristics: 1) the existence of sensors capable of allowing it to exchange data with the environment; 2) the ability to learn from experience and interact with the environment; 3) the existence of material support; 4) the ability to adapt and 5) the absence of life in the biological sense (European Union, 2017).

Among the recommendations on the constitution of a suitable registry, the formation of insurance schemes and compensation funds is the suggestion of the creation of a legal status of robots for more complex artefacts, which would then be endowed with a legal (electronic) personhood.

Electronic personhood is presented as an answer to the problems of liability in view of possible damage that could be caused by robotic artefacts. Indeed, we can discern some confusion in this form of approach: the attribution of a supposed legal personhood for robots is treated as an automatic consequence of the debate on liability. As Pagallo (2018) noted, just as we should not confuse apples with oranges, it is important to separate apples from liability and oranges from personhood. In addition to polarising the debates, it should be noted that the defence or criticism of legal personhood for robots necessarily involves an understanding of the process of conferring legal personhood on business companies and corporations. If the law does not restrict the attribution of legal personhood to human beings, how could we criticize the attribution of legal personhood to a robot ? Should we approach robotic artefacts by means of old categories, as if robots were people for the law?

THE ANDROID FALLACY AND ANTHROPOMORPHIC RHETORIC

Robots as Natural Persons

Tracing the relationship between Law and new technologies is not an easy task and, generally, this harmonisation does not occur in a simple way. This link is often made possible through the use of metaphors, which serve as an instrument for the achievement of a rhetorical effect, directly comparing different concepts. Richards and Smart (2013) explain that, when dealing with different types of robots, there are a series of competing metaphors, so choosing which ones to use generates consequences of great importance for the success or failure of an attempt to regulate Robotics.

Calo (2015) asserts that currently we are already dealing with the choice of metaphors for robots, as drones have already been equated with "aircraft", leading to severe limitations of usage. In addition, regulatory agencies in the United States have already compared surgical robots to laparoscopic surgery, which is minimally invasive, speeding up the process of approval.

A particularly seductive metaphor, not only for the law but also for other fields of study of Robotics, is to think of robots based on anthropomorphic rhetoric, as if they were people. If the imagery about robots is marked by the presence of anthropomorphic artefacts, such as the androids of films and literature, what would be the problem for the law to resort to this subtle comparison as well? To understand the risks of this rhetoric, which projects human qualities on robots with AI, we need, first, to better understand this technology.

Faced with the challenges brought by the spread of intelligent robots, which are gradually coming onto the market and consequently are becoming more and more present in people's lives, also impacting the sphere of Law, Calo (2015) presents three distinctive characteristics of robots: embodiment, emergence and social meaning. One of the main characteristics of a robot is to be physically incorporated into the world, which allows it to share the physical environment with human beings. As Mataric (2014) points out, corporeality also means perceiving other bodies and objects around it, because one of the first things a robot must internalize when programmed is how to avoid collisions, which is done with the help of sensors, physical devices that allow a robot to receive information about itself and the objects around it. In this sense, contrary to what it may seem, uncertainty is part of Robotics and arises from the fact that robots are physical mechanisms that operate in situations in which it will be difficult to know exactly their own state and that of their environment.

Materiality is not just a purely aesthetic issue. The way we think about robots (and their human operators) will also affect their design. In this context, Richards and Smart (2013) question what society expects of robots based on metaphors: are they virtual butlers, virtual pets, or virtual children? The answers chosen for these questions will affect the physical presentation of the robot and its configuration. According to Coeckelbergh (2009), ascribing responsibility to such agents is to experience, feel and perceive a form and performance. In this sense, one could speak of "virtual agency" and "virtual responsibility" to refer to "the responsibility that humans attribute to each other and to (some) non-humans based on how the other is experienced and appears to them" (Coeckelbergh, 2009).

Despite its anthropomorphic traits, Sophia, the humanoid robot, follows a simple program. On this point, the metaphor can be transmuted into a fallacy: human appearance can lead us to think of robots as people. Thus, since not all robots are androids, the illusion caused by anthropomorphism of form can be dangerous when we think of regulatory initiatives based on false assumptions about the capacity of robotic artefacts themselves.

The projection of human characteristics on robots does not depend on their form. Even when a robotic artefact has no anthropomorphic shape, people project onto these technologies human qualities such as consciousness and intelligence. As the autonomy of the system increases, making connections between the inputs (its commands) and the behaviour of the robot difficult, analogies with human beings are reinforced, which, in turn, can hinder any normative attempt, whether in terms of ethical debate, or in legal matters, such as the determination of who would be liable for possible damage caused by robotic artefacts.

The Naturalisation of Autonomy and Consciousness in Robotics and in AI

In the debate on electronic personhood, it is commonly observed that already existing legal norms would be incapable of portraying and, consequently, disciplining autonomous, intelligent robots. Since it is admitted that today's robots can perform unanticipated behaviour, we would only have to recognise their legal (electronic) personhood. This kind of reasoning has several flaws. The first is the lack of determination of the meaning of autonomy. At the same time, autonomy is confused with unpredictability of the result. Machines operated by direct human control can bring about unpredictable results. From a technological perspective, could the term "autonomy" be used in robotic applications where teleoperation, telepresence or human supervision are found at some point? Could a robot acting without constant human monitoring, but controlled at a time of need, be qualified as autonomous (Bertonili, 2013)? In this sense, the absence of specification of the term "autonomy" contributes to its own naturalization, that is, autonomy is presented as a given, as if it were a necessary consequence of the supposed intelligence of these systems.

In an attempt to dispel this imprecision, Bertolini (2013) highlights three meanings for the term autonomy when discussing robotic applications: 1) autonomy as consciousness or self-consciousness, which would lead us to the idea of free will and, consequently, to the identification of a moral agent; 2) the capacity to interact independently in the operational environment; 3) the capacity to learn.

In philosophical terms, autonomy, in a strong sense, is related to the idea that responsibility can only be attributed to a moral agent. Like subjectivity, autonomy, in that sense, is part of the philosophical discourse of modern times. Moral concepts in "modern times" have come to be shaped to recognize the subjective freedom of the individual in discerning as valid what they should do. By breaking with the paradigm of morality as obedience, Kant practically invented the concept of morality as autonomy (Schneewind, 1998). The rejection of the inequality of moral quality makes each one their own legislator, to the extent that every person would be capable of evaluating their own action, without the need for any external interference. Although the strong anthropocentric component of this idea of autonomy can be criticised, there is currently no robotic artefact that meets these described conditions, which would in principle rule out qualifying robots as autonomous agents in a strong sense. Since the law does not restrict legal personhood, as an aptitude to acquire duties and rights, to the human substrate, the ontological debate ends up losing space when confronted with more pragmatic arguments, such as the attribution of legal personhood to corporations and other business associations.

In another sense, autonomy could be understood as the ability to perform tasks without human supervision. This is autonomy in a weak sense. From the autonomous drone, to vehicles without a driver, to the robotic vacuum cleaner, one can speak, in these cases, of autonomy at various levels, even if the robotic artefact is associated with performing a certain activity due to a goal previously defined by a programmer. Although far from the idea of a strong agency concept, it is undeniable that this is an appearance of agency, which, as we have seen, has its importance. In the classic definition of Richards and Smart (2013), robotic artefacts are analysed from this sense of agency, which is not to be confused with its strong sense. In this aspect, a robot can be understood as a built system that displays, even if only apparently, a physical and mental agency, but is not alive in the biological sense, that is, it is something manufactured, that moves around the world (materiality), seems to make rational decisions about what to do (weak or apparent autonomy) and is a machine.

To avoid anthropomorphic rhetoric, Calo (2015) avoids the use of the term "autonomy" and prefers to use the term emergence. This behaviour is found in complex adaptive systems where there is a global behaviour resulting from individual interaction. Some examples can be seen in the animal world, such as the flock of birds, the school of fish and the swarm of bees, which show the creation of patterns without the existence of a central command. Emergent behaviour is a characteristic phenomenon of complex adaptive systems (Doneda et al., 2018). It is a type of global behaviour, which can result from hundreds and thousands of simple individual interactions. They create the illusion of central coordination. We speak of emergence when we observe a behaviour that is not explicitly programmed, but which results from the interaction of simple mechanisms.

The notion of emergence is associated with a holistic perspective, in which the robot's behaviour is not confused with the simple sum of its parts, creating, in some situations, the sensation that the artefact performed an unexpected, non-programmed behaviour. It is interesting to realize that surprise can depend on the subjective expectation of the expecter. Even so, even if one adopts the perspective of the programmer, there is no way to establish beforehand all the behaviours that emerge from the interaction that occurs only in a certain time and space of the execution. As Mataric (2014) points out, the fact that we cannot predict everything in advance does not mean that we cannot predict anything, such as the risks associated with the use of artefacts, such as surgical robots, in a context of a particular use. Thus, the input received by the robot continues to be determinant for the behaviour it will produce, even if the latter is unexpected.

Autonomy can also be associated with a supposed ability to learn. Could the ability of a robot to acquire and elaborate data to perform its activities be equated to real learning? There are already robotic artefacts capable of deciding independently on the course of an action without any human intervention. Could the rules that determine the action and decisions be changed by the robotic artefact itself? What does this machine learning consist of? AI systems need the ability to acquire their own knowledge by extracting patterns from raw data. This resource is known as machine learning. The learning process, which may or may not be supervised, allows the system itself to do the same task more efficiently with each attempt, thus automatically improving its experience. Among the types of learning, the outstanding one today is deep learning, which attains great power and flexibility in the attempt to represent the outside world with an aligned hierarchy of concepts, allowing the classification of images, speech recognition and object detection, among other uses.

As Goodfellow et al. (2016) point out, the first deep learning algorithms we recognize today were thought of as computational models of biological learning, that is, models of how learning happens or can happen in the brain. Deep learning is closely associated with the architecture of artificial neural networks. Here it is noted that anthropomorphism is not a unique characteristic of Robotics. AI has also been historically conceptualised in anthropomorphic terms. As Watson (2019) points out, besides the fact that people always talk about machines that think and learn, the name itself (artificial intelligence) challenges us to repeatedly compare human ways of reasoning with algorithms. In the same way as with legal entities, it is not always clear whether this language is used in a literal or metaphorical sense.

The anthropomorphic metaphor conceals functional aspects of artificial intelligence, so that this rhetoric, which mimics human qualities and attributes, may compromise the response to the complex ethical challenges posed by emerging technologies. In fact, it is a mistake to suppose that these algorithms can be confused with human intelligence, since, although they surpass human intelligence in certain aspects, they also fall short in others (Watson, 2019). Even though one cannot criticize simple inspiration in human models for the development of artificial intelligence, it is always important to be careful when differences are erased and one begins to think of metaphors and analogies in their literal sense. Consequently, when thinking about any attempt to discipline or regulate Robotics, it is fundamental not to confuse the existence of real autonomy or agency with the sensation of autonomy or agency. Unfortunately, the confusion between the supposed agency of the artefacts and the sensation provoked by the emerging technology leads to a naturalization of the autonomy itself, as if every robot with AI necessarily was, as happens with human beings, making a decision in a specific and independent way.

Social Robots, Vulnerability and Social Valence

It is important to separate certain issues that are confused in the debates about the legal personhood of robots. Anthropomorphism does not depend on the beliefs people may have about the ontological nature of artefacts. Even acknowledging that current questioning about the status of "intelligent" robots may impact on how people reflect and relate to these artefacts, the debates about the supposed agency of robots, or about the technical possibility of developing a complex artificial intelligence system, called strong AI, may not condition people's willingness to continue to explain the behaviour of a robotic artefact based on the assignment of mental states. This happens on account of the particular social valence of this technology.

Moreover, social meaning (or social valence) relates to the fact that humans show greater social commitment and provide different stimuli when dealing with robots compared to other goods. This characteristic can be linked to embodiment, since the physical embodiment of the robot tends to make a person treat that moving object as if it were alive. This is even more observable when the robot has anthropomorphic characteristics, since the resemblance to the human body causes people to start projecting emotions, feelings of pleasure, pain and care, as well as desires to constitute relationships. Balkin (2015) understands that the projection of human emotions on inanimate objects is not a recent phenomenon in human history, but when applied to robots, it entails numerous consequences.

Calo (2015) lists some consequences that can be generated by social valence, amongst which Balkin (2015) highlights four: 1) the more anthropomorphic a robot is, the more people blame the robot, rather than the person who uses it; 2) the presence of robots in a surveillance system increases the subjective feeling that someone is being watched; 3) humans take greater risks to preserve the integrity of anthropomorphic robots than for things designated as tools; and 4) humans may suffer distinct emotional damage from the loss of robotic fellows.

Robotics is no longer restricted to the factory and the laboratory. So-called social robots are designed precisely to interact with humans in uncontrolled environments. To this end, studies and projects have been intensified to develop artefacts capable of interacting with people as naturally as possible. Social robots are characterized by the possibility, albeit apparent, to transmit emotions, encourage and form social relationships, demonstrate personality, use natural clues of communication and interact socially with people. There is already a particular field of study called human-robot interaction (HRI), which seeks, based on social valence, to replicate in robotic artefacts a variety of cues and markers present in human communication, such as facial expressions and even language.

Along with social robots, assistive and rehabilitation Robotics also stand out. Pearl, the Nursebot, is a prototype of a personal mobile robotic assistant that can recognize speech, accompany patients and communicate via touch screen. Designed at Carnegie Mellon University, the nurse robot is being prepared to remind people to take their medicine and help them move around in old people's homes. Rehabilitation robots were initially designed to assist in the movement of patients in recovery. Assistive Robotics has always had a wide reach, including rehabilitation robots, wheelchair robots, companion robots and manipulative arms. We can also speak of a Socially Assisted Robotics, a term used to describe artefacts whose central focus, instead of physical contact, is some form of social interaction. Robots are already used to help stroke (CVA) patients to do their exercises, to assist the elderly and to care for and educate children and adolescents, especially in cases of specific conditions, as has been advocated in situations of autism.

According to Sharkey and Sharkey (2008), there are several ethical problems related to the use of social robots by people in vulnerable situations. With regard to the elderly, the following are noteworthy: 1) potential reduction in human contact; 2) increased sense of objectification and loss of control; 3) loss of privacy; 4) loss of personal freedom; 5) deceit and infantilisation; 6) uncertainty regarding the circumstances in which the elderly can and should have permission to control robots. For Sparrow and Sparrow (2006), the use of social robots with the elderly reveals a serious ethical problem, as it is based, mainly in the case of anthropomorphic artefacts, on the illusion of genuine social interaction. Even in the case of relatively simple assistive robots, introduced in old people's homes to monitor their behaviour, one can speak of a technology that decisively affects the choices of these people, which can result in authoritarian Robotics.

When we think of robots as if they were people, we envisage for the artefact a degree of agency and autonomy that is not simply exaggerated, it is actually a transference, in which we lose part of our own autonomy. The proposal of an electronic personhood does nothing to help deal with this problem. It may, in fact, aggravate it, since, even if it is restricted to Law, legal personhood reinforces the concealed equivalence that is symbolically projected towards other fields. But if we move the artefacts away from the idea of natural person, would we not run the risk of abandoning our own ethics in these interactions, as can be seen, for example, with the advance of sexual robots that reproduce misogynistic stereotypes present in society? The social valence of robots shows us exactly the opposite, that is, that ethics can and must precede the definition of the nature of these technologies, by the simple fact that we are human beings, "with autonomy and moral rules, dealing with these ontologically indefinite artefacts" (Cortese, 2018).

A virtue ethics approach can thus offer an interesting way of dealing with the problems generated by the interaction between humans and social robots. To avoid the risk of an individualist solution, Coeckelbergh (2010, 2020) highlights the importance of thinking about a relational and socially oriented ethics of virtue, that is, "virtue in its history and in its concrete bodily performances" (Coeckelbergh, 2020).

The "individualist solutions", which also mark the philosophical discourse of modernity, have also been transposed to legal discourse. The emphasis placed on the subjective centre of abstract imputation stems from the transposition of an illusion: the individual-subject of law with all his attributes would be capable of shaping the whole juridical system (Alcaro, 1976). While, on the philosophical level, the philosophy of conscience favoured the immediacy of subjective experience over discursive mediation (Habermas, 2007), on the juridical level, processes of social interaction, such as the union of persons around a certain initiative, also came to be portrayed by the interposition of a transcendental subjectivity: the legal person.

ELECTRONIC PERSONS AS LEGAL ENTITIES

Legal Entity and Calculation With Concepts

The main argument for the defence of electronic personhood is associated with a pragmatic or functional analysis of legal personhood. In the verbalized legal world, the term "legal person" refers to an autonomous centre of legal relations. If legal personhood is already dissociated from the human substrate, there would be no way to deny personhood to robots due to the non-existence of any human characteristic in these artefacts. In that narrative, the legal person is presented as if there were no problems in the process of attributing legal personhood to companies. The analogy with the legal person requires, however, an understanding of the function of this term in legal grammar.

The philosophical discourse of modernity is not structured only in subjectivity. The rationalization that crystallizes around the organization of the capitalist enterprise and the bureaucratic apparatus of the state also appears as an essential characteristic of those "new times", with the institutionalization of economic and administrative action with regard to the aims. Law is also going through a process of rationalization, the central idea of which is the differentiation and institutionalisation of autonomous social systems, thought of as machines, since they are founded on themselves and governed by a particular procedural reason. The consolidation of this formal law is not limited to the external foresight of the administration of justice or to the separation of powers, but also requires an internal, predictable control, embodied in the idea that it is "calculated with concepts", as in mathematics.

The term legal person was perfectly suited to the context of formal Law internally controllable by means of abstract concepts. Even today, when we perceive that this pretension of a legal machine has always been illusory and Law is incalculable, as Irti (2018) pointed out, we can also see that the legal person retains, to a certain extent, its original inspiration: calculation mediated by concepts.

Functions and Illusions of the Legal Person

According to Solaiman (2017), being a legal person entails the ability to exercise rights and to perform duties. For Bryson et al. (2017), there are three issues related to legal personality that directly interest the debate on electronic personality. First, legal personality is a fiction. Legal personality is not necessarily correlated with an ethical notion of moral personhood. Second, legal personality is divisible. A legal system might treat differently legal entities in respect of some rights and some obligations. Third, the rights and obligations that a legal person may have as a matter of law may not match those it has as a matter of fact (Bryson, 2018). Even agreeing with the points presented, we believe that the heuristic function of the term legal person has a decisive role in the analysis of the proposal to create an electronic personality.

The legal person represents a mental shortcut, a trigger that facilitates access to a set of complex situations. The acts performed by shareholders and directors are unified around abstract subjectivity, and there is no need, in each situation, to refer to the whole set of people who are contemplated by the legal entity's particular framework. In this sense, it is important to perceive the heuristic function of the term legal person, that is, a mental shortcut that enables, with simplified information, rapid judgements.

As a mental shortcut, legal personhood allows the allocation of the patrimony in autonomous centres, different from the complex of legal relations of each partner. The creation of the new subject (legal person) facilitates the understanding of the separation of assets according to a particular purpose. This, however, creates the illusion that patrimonial segregation is dependent on legal personhood, as if patrimonial autonomy could only be explained with the mediation of the legal person. In addition to the simplification of the complex of relationships and the autonomous allocation of assets, recourse to corporation personhood also allows access to a model of private imputation of acts practiced by shareholders and directors and, at the same time, gives stability to the model of coordination that develops within the legal person.

In the debate on electronic personhood, the process of conferring legal personhood on companies is presented as a model that would justify the recognition of legal personality for robots with artificial intelligence, as argued, for example, by Turner (2018), who even maintains that possible abuses, such as the lack of accountability of programmers and engineers, could be fought by disregarding legal personhood ("piercing the corporate veil"). This type of argument demonstrates how the analogy with corporate law is mobilized without, for this purpose, pointing out the problems present in the model of the corporate personality.

As Galgano (2010) had already reported in Italian law, there are several disadvantages in the process of conferring legal personhood on companies, which are not, to this day, properly measured. Galgano (2010) pointed out that the term legal person was used, both by courts and lawyers, as if there was a single entity to be protected behind the label of the legal person. This form of approach generated a serious problem: unitary treatment. Besides distorting the function of the institute, it masked the diversity of phenomena that articulated around that term. Similarly, Ferro-Luzzi (2001) demonstrated how the idea of activity, fundamental to the understanding of the term enterprise, was mistakenly absorbed by the notion of abstract subjectivity, which, in turn, compromised the very regulation of the business phenomenon by the law. According to the Italian author, the concept of activity depends on a new legal grammar, which reveals itself capable of culturally disassociating the action from the figure of the abstract subject that has rights and duties.

The model of the corporate personality has also contributed to an improper understanding of the limitation of the shareholder's liability by concealing the unequal transfer of entrepreneurial risk to third parties. If, on the one hand, there are creditors who can protect their own interests by renegotiating the risk with the company, as happens with a financial institution; there are, on the other hand, creditors who are unable to do so, as is sometimes seen with victims of environmental damage, such as those affected by mining. The prevalence of the abstract model of subjectivity has given rise to a unitary reading of patrimonial autonomy itself and, consequently, of the limitation of responsibility, which are indifferent to the different credits.

If the electronic personhood has been conceived according to the problems generated by the need to be accountable for possible damages, it should be remembered that there is a mismatch between the legal format of the isolated corporation and the economic protagonism of the multinational enterprise groups. This is an internal contradiction of Law, materialized in the paradoxical tension between legal diversity and economic unity. To minimize this problem, Law has sought a new grammar, coming closer to the figure of control and direction, breaking with the model of an abstract subject as the central point in the process of accountability.

The creation of an electronic personhood may end up repeating the same problems. Instead of recognizing the peculiarities of the different areas of operation of robots, these different relationships are unified in a single legal model, based exclusively on the figure of an abstract subject. This is a frequent mistake when the law tries to approach new technologies. Instead of their ownership, the artefacts are in fact determined by their specific destinies. Thus, they do not include abstract generalizations and unitary reductions, regardless of their various uses. Is it possible to compare the problems caused by the use of Robotics in medicine with the use of drones for military and security purposes? Similarly, the use of social robots with vulnerable people raises specific ethical problems, which cannot be compared with the use of Robotics for the transport of goods and people.

Accountability focused on the personhood of this new subject, supported by a still debatable concept of autonomy, may conceal those who are truly responsible for the damage and for the development of the artefacts, transferring the risks of the activity carried out by programmers and computer engineers to third parties who share the same spaces with the robots. Contrary to what Turner (2018) states, "piercing the corporate veil doctrine" (disregard of legal entity) does not represent an adequate instrument to remedy these problems, but represents, in fact, a technique that is the main manifestation of the unitarianism that marks the whole discourse of the legal person. There can be seen in the European Parliament's particular Resolution with recommendations on Civil Law on Robotics, confusion between the attribution of personhood and the separation of patrimony. The creation of a specific fund for any damage caused does not depend on the creation of a new subject, since the legal person, even if associated with patrimonial autonomy, does not have a monopoly on the disposition of property. Nor does criticism of the personification make the disposition of property the main solution to the problem. It is fundamental to come up with differentiated liability mechanisms, sensitive to the different uses of robotic artefacts and the diverse types of damage that may possibly be caused.

On April 21, 2021, the European Commission presented the Proposal for a Regulation on Artificial Intelligence, which seeks to establish a uniform legal framework for the development, commercialization, and use of artificial intelligence within the scope of the European Union. The current proposal moved away from the creation of an electronic legal personality. The text relies on a risk-based approach, which modulates the content of standards according to the intensity of risks created by AI systems.

Taking Metaphors Seriously: New Subjects and the "Imitation Game"

The proposal to create an electronic personhood is part of a wider debate: the recognition of new subjectivities and, consequently, new legal actors (Gellers, 2020). Teubner (2006) recalls that in 1,522 rats were submitted to a trial in the ecclesiastical court of Autun. The methodological individualism that has informed legal personhood since modern times has prevented the recognition of animal rights. Influenced by the process of rationalization of science and nature, the number of actors in the legal world was, as the German author maintains, drastically reduced by a development of the philosophical discourse of modernity. In dialogue with Luhmann's Theory of Systems and with Latour's sociology, Teubner (2006) rejects the anthropocentrism that underlies the psychological and sociological analysis of an intentional action in which the only plausible actor is the human individual.

In 2017, a river in New Zealand was given legal personhood. In the same year, in India, a court recognized the legal personhood of the rivers Ganges and Yamuna. Unlike the Indian case and the New Zealand case, the Constitution of Ecuador made a more daring proposal. The projection of the rights of nature was presented as a way of trying to move from an anthropocentrism to a biocentrism based on the idea of good living. This openness to new forms of subjectivity has the merit of trying to dissociate oneself from the individualistic model that underlies both the natural person and the legal person. But might it be possible to combat anthropocentrism by making use of an instrument such as the legal personhood, the main representative of methodological individualism in legal grammar? Even if these initiatives are of great importance, in a symbolic and cultural dimension, by recognising the wisdom of traditional and indigenous populations with a new cosmovision, the new personalities may end up imprisoned in an old grammar still inspired by an anthropocentric model, such as the ideas of subjective rights and individual ownership. The same can happen with the supposed electronic personhood. Even if the association with the dichotomy natural person and legal person is avoided, the new subjects are articulated by means of old models, which reinforce the already classic subjective modulation of legal discourse.

In the lesson of Rodotà (2015), the problem lies in the perspective of the very idea of an abstract subject that informs any process of attribution of legal personhood. This construction allowed the juridical discourse to formally liberate the person, artificially detaching him or her from his or her economic, social and natural conditions. As a response to the contempt for the concrete, we note the attempt to reconnect the person, in a material sense, to his or her context, with the reinvention of the person, now socio-environmentally situated and embodied.

The pitfall of the metaphor of the abstract subject is precisely that it tends to merge person and juridical subjectivity by not demonstrating the differences and thus hiding them. In Serick's classic study (1958), there is reference to the teratological case People's Pleasure Park Co. v. Rohleder, in which a Virginia court in 1908 asked itself what the colour of the legal person would be when faced with the following question: whether a society, as an autonomous centre of juridical relations, could be constrained by the racist laws of the state which prohibited blacks from acquiring land. In Germany, with the rise of Nazism, the courts also had to examine whether the anti-Semitic laws could be applied to companies controlled by Jews (Serick, 1958).

In the case Santa Clara County v. Southern Pacific Railroad, the term "person", provided for in the 14th amendment of the US Constitution, was also associated with a corporation, which could be seen as an example of a subject for Law (Hall, 2005). In 2014, in a controversial decision, the US Supreme Court resorted to the argument that an entrepreneurial society, Hobby Lobby, could invoke religious freedom in order not to collaborate with the payment of a health plan that would allow employees access to emergency contraceptive drugs, with high doses of oestrogen, popularly known as morning-after pills.

The accommodation of the religious freedom of a for-profit business society comes up against an important point, however: thousands of women employed by Hobby Lobby may not share the same belief as the main shareholders in the company. In view of this situation, did the court decide to protect the legal position of the company's controlling shareholders to the detriment of the private autonomy of the female employees? For Judge Ginsburg, the casting vote at the time of the trial, there was no doubt: the choice to extend religious freedom to a profit-making organisation generated a serious imbalance within the company by favouring the belief of the controllers over the protection of the rights of women working in the company in question.

In the debate on the rights of the personhood of legal persons and on the moral damage to legal persons in Brazil, there is a sometimes problematic approach between natural persons and legal persons. This equalisation may, as already highlighted, ignore the diversity of interests that justified the personification of the human being in relation to the embodiment of companies, foundations and associations. Just as it is important to criticize the disguised fusion between person and legal person, we should also separate person and legal personhood and recognize that the expansion of new subjects refers only to the latter, to juridical subjectivity.

In this context of new subjectivities, what should be done? Albeit controversial, the very origin of the term legal personhood, derived from the term *persona*, is associated with a metaphor, the mask used in theatre, allowing the actor to impose his voice. Despite this remote use, people still believe today in the illusory possibility that metaphors, even those already incorporated within legal grammar, can be prohibited. Italian nominalism, recognising that the legal person would represent a linguistic instrument, almost suggested its end, thus underestimating the power and function of metaphors. Even if there is no way to eliminate them, it will always be possible to monitor their normative use, reporting, in specific situations, the abuses related to the use of metaphors and analogies in a literal sense.

As Turner (2018), one of the enthusiasts of the attribution of legal personhood to robots, points out, the accreditation of electronic personhood to robots in the United States or the European Union is likely to influence other legal decisions. The electronic personhood may thus be adopted by countries that traditionally import legal models, as is the case of Brazil, whose model of legal personhood for natural persons has never been fully achieved. Political, economic and social challenges have prevented, and still prevent, the construction of a complete citizenship in several peripheral countries. Although influenced by the philosophical discourse of modernity, the adoption of legal models in Brazil has occurred, in various situations, in a particular and partial way, as in a real game of imitation, an incomplete and untimely simulacrum of never-realised expectations. We cannot move on to new subjectivities without confronting old promises, such as the problems of subjects whose human rights have not yet been achieved, at the risk of confusing people and legal entities. Perhaps robots with artificial intelligence can wait for their controversial rights. Perhaps the only task, no less important, left for us to carry out is that of adjusting subjects, putting back on the masks and taking the metaphors seriously, that is, continuing to report the non-problematised convergence between the contemplated metaphor and the disguised comparison.

CONCLUSION

The sentence "the robots are coming", which has already become a cliché, does not accurately portray the evolution of this technology. If robots, in fact, have already arrived, what is this so loudly proclaimed Robotics revolution? Robotic artefacts, in contrast to what used to happen, are increasingly integrated into the same environments as human beings, which, in turn, can have great impacts, not yet fully measured, as can be seen in the use of these technologies in medical care and care for the elderly and children. The imaginary about robots is intensely marked by the association with anthropomorphic artefacts, such as androids, which appear in films and literature. A particularly dangerous metaphor for the law is to yield to this symbolism, projecting autonomy, consciousness and other human attributes into robotic artefacts. Often the different concepts, originally fused around the metaphor, disappear, so that differences are erased and metaphors and analogies come to life, coming to be thought of in their literal sense.

The discussion about the ontological foundations that separate people and robots has been seen to be insufficient to remove the defence of legal personhood from robotic artefacts with artificial intelligence. If the law confers legal personhood on assets intended for certain purposes, such as foundations, there can be no doubt that the aptitude to acquire rights and duties is not exclusively one of human beings. In fact, we note the prevalence of a pragmatic or functional line of the electronic personhood, which, by distancing itself from the philosophical debate centred on ontological analyses, seeks to base itself mainly on the model of the corporate legal personality. This change of focus, with robots as legal persons, also involves problems, which in most cases are neglected even by critics of the electronic personhood. This is mainly on account of the incorrect understanding of the reasons present in the process of embodiment of companies and the particular role of the term “legal person” in the grammar of Law.

In Jorge Luis Borges’ fictional essay, the substitution of the anthropomorphic metaphor by a sphere inspired several thinkers, until it became a labyrinth and an abyss for Pascal, who, feeling the incessant weight of the physical world, adjusted his metaphor, going on to claim that “nature is an infinite sphere, whose centre is everywhere and its circumference nowhere”. Blaise Pascal, whose studies were fundamental for computing, was also known for his wager as to the infinite. In this single player game, we can reflect ethically on the existence of the indefinite, even if it is rationally inaccessible. In the same way, we do not need to wait for ontological definitions or these robotic artefacts to definitively become part of people’s everyday lives to question ethical problems related to this process. Should we be concerned about social robots? What are the main risks associated with the so-called Socially Assistive Robotics? If, on the one hand, the electronic personhood contributes very little to the problems generated by the not at all metaphorical approximation between robots and humans; on the other hand it reinforces dangerously the connection, not always questioned, between anthropomorphic rhetoric and concealed imitation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Alcaro, F. (1976). *Riflessioni critice intorno alla soggettività giuridicha: significato di una evoluzione*. Milan: Giuffrè.

Balkin, Jack. (2015). *The Path of Robotics Law*, 06. Berkeley: California Law Review Circuit, 45–60., v.

Bertolini, A. (2013). Robots as Products: The Case for a Realistic Analysis of Robotic Applications and Liability Rules. *L. Innovation Techn.* 2, 214–247. doi:10.5235/17579961.5.2.214

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: The Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25 (3), 273–291. doi:10.1007/s10506-017-9214-9

Bryson, J. J. (2018). Patiency Is Not a Virtue: the Design of Intelligent Systems and Systems of Ethics. *Ethics Inf. Technol.* 20, 15–26. doi:10.1007/s10676-018-9448-6

Calo, Ryan. (2015). Robotics and the Lessons of Cyberlaw. *California Law Review*. 1033. 513–563. v. n.

Coeckelbergh, M. (2020). How to Use Virtue Ethics for Thinking about the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance. *Int. J. Soc. Robotics* 13, 31–40. doi:10.1007/s12369-020-00707-z

Coeckelbergh, M. (2010). Robot Rights? towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12 (3), 209–221. doi:10.1007/s10676-010-9235-5

Coeckelbergh, M. (2021). Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach. *Int. J. Soc. Robotics* 1. doi:10.1007/s12369-021-00770-0

Coeckelbergh, M. (2009). Virtual Moral agency, Virtual Moral Responsibility: on the Moral Significance of the Appearance, Perception, and Performance of Artificial Agents. *AI Soc.* 24, 181–189. doi:10.1007/s00146-009-0208-3

Cortese, J. (2018). Interação, indistinguibilidade e alteridade na Inteligência Artificial, 17. *Teccogs: Revista Digital de Tecnologias Cognitivas*. São Paulo, Brazil: TIDD | PUC-SP, 95–112.

Darling, K. (2016). in *Extending Legal protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects*. Editors R. Law., R. Calo, and A. M. Froomkin (MA: Edward Elgar Publishing), 213–231.

Doneda, D., et al. (2018). Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal. *Pensar - Revista de Ciências Jurídicas. V.* 23–17. doi:10.5020/2317-2150.2018.8257

Galgano, F. (2010). *Trattato di Diritto Civile, I.* (Milan: CEDAM).

Ferro-Luzzi, P. (2001). *I* contratti associativi. Milan: Giuffrè, 1–16.

Gellers, J. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. Abingdon: Routledge.

Goodfellow, I., Bengio, Y., and Courvile, A. (2016). *Deep Learning*. Cambridge: MIT PRESS.

Gunkel, D. J. (2018b). *Robot Rights*. Cambridge, Massachusetts: The MIT Press.

Gunkel, D. J. (2018a). The Other Question: Can and Should Robots Have Rights? *Ethics Inf. Technol.* 20 (2), 87–99. doi:10.1007/s10676-017-9442-4

Habermas, J. (2007). *The Philosophical Discourse Of Modernity. Twelve Lectures. Fredrick Lawrence (Translator)*. Cambridge: MIT PRESS.

Hall, K. L. (2005). *The Oxford Companion to the Supreme Court of the United States*. Oxford: Oxford University Press.

Irti, N. (2018). *Un Diritto Incalcolabile*. Torino: Giappichelli.

Leenes, R. (2011). Framing Techno-Regulation: An Exploration of State and Non-state Regulation by Technology. *Legisprudence* 52, 143–169. doi:10.5235/175214611797885675

Mataric, M. J. (2014). *Introdução à robótica. Humberto Ferasoli Filho, José Reinaldo Silva e Silas Franco((Translator)*. São Paulo: Editora Unesp.

Pagallo, U. (2018). Vital, Sophia, and Co.-The Quest for the Legal Personhood of Robots. *Information* 9, 230. doi:10.3390/info9090230

Parviainen, J., and Coeckelbergh, M. (2020). The Political Choreography of the Sophia Robot: Beyond Robot Rights and Citizenship to Political Performances for the Social Robotics Market. *AI & SOCIETY* 36, 715–724. doi:10.1007/s00146-020-01104-w

Richards, N. M., and Smart, W. (2013). *How Should the Law Think about Robots?* Available at SSRN 2263363.

Rodotà, S. (2015). *Il diritto di avere diritto*. Roma: Editori Laterza.

Schneewind, J. B. (1998). *The Invention of Autonomy*. Cambridge: Cambridge University Press.

Serick, R. (1958). *Apariencia y realidad en las sociedades mercantiles: el abuso de derecho por medio de la persona jurídicaJosé Puig Brutau*. Translator. Barcelona: Ediciones Ariel.

Sharkey, A., and Sharkey, N. (2010). Granny and the Robots: Ethical Issues in Robot Care for the Elderly. *Ethics Inf. Technol.* 141, 27–40. doi:10.1007/s10676-010-9234-6

Solaiman, S. M. (2017). Legal Personality of Robots, Corporations, Idols and Chimpanzees: a Quest for Legitimacy. *Artif. Intell. L.* 25 (2), 155–179. doi:10.1007/s10506-016-9192-3

Sparrow, R., and Sparrow, L. (2006). The Hands of Machines? The future of Aged Care. *Minds Machines* 16, 141–161. doi:10.1007/s11023-006-9030-6

Teubner, G. (2006). Rights of Non-humans? Electronic Agents and Animals as New Actors. *J. L. Soc.* 33, 497–521.

Turner, J. (2018). *Robot Rules: Regulating Artificial Intelligence*. Berlin, Germany: Springer.

Watson, D. (2019). *The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence*. Minds & Machines. doi:10.1007/s11023-019-09506-6

WINNER (1985). “Do Artifacts Have Politics,” in *The Social Shaping of Technology* (Philadelphia: Open University Press).

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 03 January 2022
doi: 10.3389/frobt.2021.791527

Empathizing and Sympathizing With Robots: Implications for Moral Standing

*Oliver Santiago Quick**

Research Unit for Robophilosophy and Integrative Social Robotics, Aarhus University, Aarhus, Denmark

This paper discusses the ethical nature of empathetic and sympathetic engagement with social robots, ultimately arguing that an entity which is engaged with through empathy or sympathy is engaged with as an "experiencing Other" and is as such due at least "minimal" moral consideration. Additionally, it is argued that extant HRI research often fails to recognize the complexity of empathy and sympathy, such that the two concepts are frequently treated as synonymous. The arguments for these claims occur in two steps. First, it is argued that there are at least three understandings of empathy, such that particular care is needed when researching "empathy" in human-robot interactions. The phenomenological approach to empathy—perhaps the least utilized of the three discussed understandings—is the approach with the most direct implications for moral standing. Furthermore, because "empathy" and "sympathy" are often conflated, a novel account of sympathy which makes clear the difference between the two concepts is presented, and the importance for these distinctions is argued for. In the second step, the phenomenological insights presented before regarding the nature of empathy are applied to the problem of robot moral standing to argue that empathetic and sympathetic engagement with an entity constitute an ethical engagement with it. The paper concludes by offering several potential research questions that result from the phenomenological analysis of empathy in human-robot interactions.

Keywords: sympathy, empathy, HRI, moral status, phenomenology, social robot

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Janina Loh,
Stiftung Liebenau, Germany
Luisa Damiano,
Università IULM, Italy

****Correspondence:***
Oliver Santiago Quick
osquick@btech.au.dk

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

Received: *08 October 2021*
Accepted: *29 November 2021*
Published: *03 January 2022*

Citation:
Quick OS (2022) Empathizing and Sympathizing With Robots: Implications for Moral Standing. Front. Robot. AI 8:791527. doi: 10.3389/frobt.2021.791527

1 INTRODUCTION

Sympathetic and empathetic robots have become an increasingly popular topic of research within HRI. While a number of experiments have suggested that humans can feel sympathy or empathy for social robots (Riek et al., 2009; Leite et al., 2013; Rosenthal-von der Pütten et al., 2014; Leite, 2015; Ceh and Vanman, 2018; Menne and Schwab, 2018), the theoretical foundations of both the empathy and sympathy concepts, as well as their connections to ascriptions of moral standing, have been underexamined within the field of HRI. This paper will draw on philosophical, sociological, and psychological research to argue that not only are the concepts (and associated phenomena) of sympathy and empathy distinct, but that the tendency to employ one or both of these concepts without sufficiently clarifying in what sense they are intended has acted as a limiting factor on the progress of HRI research investigating these phenomena. To arrive at unified terminological standards is not only of importance for the comparability of HRI studies, however; as I shall argue here, it is also directly relevant for empirical and conceptual-normative research on the moral standing of robots.

I proceed in two steps. First, I will discuss three broad notions of empathy which researchers should have in mind when employing the concept, as well as offer a novel definition of sympathy that makes clear the distinction between empathy and sympathy and the connections of both phenomena to ascriptions of moral standing. Section two will briefly present the empathy and sympathy concepts, as well as discuss why the distinction matters and consider how the terms have been used within extant HRI research, while placing an emphasis on the valuable insights from phenomenological understandings of empathy—which have been insufficiently considered—and on the important empathy-sympathy distinction. The section three will turn to an analysis of the import sympathy and empathy can have on the moral standing of social robots. I will argue there that a phenomenological understanding of empathy suggests that empathetic or sympathetic engagement with a robot already constitutes an ethical engagement (i.e., engagement with the robot as one which possesses at least "minimal" moral standing). The approach to robot moral standing offered here is similar to, yet distinct from, the relational approaches to robot moral standing that have been offered by David Gunkel and Mark Coeckelbergh (Coeckelbergh, 2012; Gunkel, 2012; Gunkel, 2018a; Coeckelbergh, 2018), and is based primarily on the phenomenological understanding of empathy offered by Edith Stein (1964) and Max Scheler and Health (1923)).

2 EMPATHY AND SYMPATHY

The term "empathy" has only existed in English for little over a century (Stueber, 2019), but the conceptual origins can be traced back at least to the 17th century discussion of "sympathy," where philosophers David Hume (Hume, 1740) and Adam Smith (Smith, 1759) leveraged the concept to explain a range of phenomena in human-human interactions that have since been further differentiated. While "empathy" has become increasingly popular as a broad label for "knowing what an Other feels" or "feeling what an Other feels," the term "sympathy" has generally become understood as "feeling bad for an Other," Even with these vague folk definitions of the two concepts, two issues with the usage of the terms in HRI research become immediately apparent. For instance, it becomes clear that there are at least two senses of "empathy," and that sympathizing and empathizing are not the same thing. Additionally, and perhaps stemming from these conceptual issues, there is a lack of sufficient conceptual care in relating empathy and sympathy to ascriptions of moral status, as will be elaborated in what follows. In this section, I will begin by discussing the two senses of empathy already suggested, as well as a third, more "basic" sense, before presenting a sympathy definition that captures the distinction between empathy (in all three senses) and sympathy.

2.1 Empathy: Cognitive, Affective, and Phenomenological Understandings

The first form of empathy, "knowing what an Other feels," is often discussed under the label of "mind-reading" (Goldman, 2006; Singer, 2006), "mentalizing" (Singer, 2006) or "cognitive empathy" (Stephan, 2015; Bloom, 2018). Cognitive empathy can be understood as a process by which we are attribute mental or affective states to an Other, but do not "share" in these states or feel them ourselves. For instance, in seeing a stranger crying, I might infer that they are sad—if so, I have cognitively empathized. I need not feel sad myself in order to reach this conclusion, nor need I care about the sadness of the stranger. Such inference-based empathizing can be understood, broadly, as falling under the "theory of mind theory" (Carruthers and Smith, 1996) understanding of how empathy occurs. On the other hand, I might also simulate, or use my imagination, to attribute the sadness in one of two possible ways. Firstly, I might imagine what would make me cry in a public setting and decide that what the crying stranger is experiencing it is most likely sadness. Secondly, if the person is someone I know well rather than a stranger, I might imagine what would make *him* cry in a public setting (i.e., by taking into account information about his attitudes, beliefs, etc.,). In either case, as in inference-based cognitive empathy, I will not feel sad myself, nor need I care about the Other for the empathy to succeed.

Indeed, it is constitutive of cognitive empathy that I do *not* feel what the Other feels, for in the case where I "share" the affect of the Other (sadness, in this case), I am actually *affectively* empathizing. Like cognitive empathy, affective empathy is typically understood as relying on either inferences or simulations, but with the addition that one experiences an affective state similar to that of the Other. For instance, my understanding and "sharing" of a rock-climber's fear can arise through my connecting aspects of her situation to affective memories of my own (Adams, 2001). Alternatively, this can also occur through imagining myself in the climber's situation (Ravenscroft, 1998), or supposing that the climber is not a stranger, through a simulation of what I believe she is likely to be experiencing. While affective and cognitive empathy are clearly distinguishable by the inclusion or exclusion of "state-matching," current HRI research tends to employ the term "empathy" without defining the term or in such a way that the boundaries between affective empathy, cognitive empathy, and sympathy become blurred.

For instance, consider a 2018 study by Ceh and Vanman, where "empathy" was measured with the two response items "I think this scenario is sad" and "I would have sympathy for someone in this situation" (Ceh and Vanman, 2018, p. 11). Believing a scenario to be sad is not the same as empathizing with a particular social agent. Likewise, sympathizing with someone goes above and beyond empathizing with them, as will be argued in the following section. Similarly, a 2009 study by Riek et al. investigating "empathy" for robots with distinct degrees of human-likeness measured the "empathy" of participants for the robots in terms of sympathy: "After each of the clips, we asked respondents a single question, 'How sorry do you feel for the protagonist?'" (Riek et al., 2009, p. 4). When they compared the results of this question to their baseline measurements of dispositional empathy—which was measured via the Empathy Quotient (EQ) (Baron-Cohen and Wheelwright, 2004)—the researchers found that higher scores on the EQ did *not* predict higher "empathy" as measured by their single question. The lack of confirmation of their hypothesis is not surprising given the way they chose to measure empathy; the EQ is largely directed at perspective taking (e.g., "I am good at predicting how someone will feel" (Baron-Cohen and Wheelwright, 2004, p. 172)) and social intelligence (e.g., "I can easily tell if someone wants to enter a conversation" (ibid. 171)). However, despite the majority of the questions on the EQ targeting "empathy," the developers of the metric actually intentionally included aspects of sympathy (e.g., pity, compassion, and concern), simply because they see sympathy "as a clear instance of the affective component of empathy," which includes a motivation to help (Baron-Cohen and Wheelwright, 2004, p. 164).

This indicates that the tendency to treat empathy and sympathy as interchangeable is not limited to HRI but represents a much larger trend, which has simply been carried over. Indeed, there are some accounts of "empathy" which employ the term in a very broad sense, encompassing affective and cognitive empathy, sympathy, compassion, emotional contagion, and a variety of other interpersonal phenomena (e.g., Preston and de Waal, 2002). The desire to adopt a definition of empathy that encompasses all of these related interpersonal phenomena is understandable, of course. Indeed, as Frederique de Vignemont and Tania Singer suggested, "There are probably nearly as many definitions of empathy as people working on the topic" (de Vignemont and Singer, 2006, p. 435). Unfortunately, such approaches to defining empathy directly conflict with what was meant by "empathy" when the term was first coined, as well as what is frequently meant by terms such as "compassion" and "sympathy." The distinction between cognitive empathy ("mentalizing") and affective empathy, for instance, is not merely a matter of terminology, but also of physiology (Singer, 2006). From a phenomenological perspective, it is also clear that a phenomenon such as emotional contagion, for instance, is explicitly *not* "empathy" (Scheler and Health., 1923; Stein, 1964; Zahavi, 2014).

The affective/cognitive empathy distinction, and the distinction between sympathy and empathy, are perhaps still underutilized, but have begun to receive attention within HRI (Asada, 2015; Stephan, 2015; Quick, 2020). However, a third *phenomenological* understanding of empathy has been largely overlooked within HRI. As I will argue, this sense of empathy, which can be more or less understood in terms of what Alvin Goldman has called "low-level mindreading" (Goldman, 2006) or Karsten Stueber calls "basic empathy" (Stueber, 2006) is perhaps the most important for the question of moral standing. Basic empathy, as opposed to "complex empathy" (Hollan, 2012) (i.e., affective and cognitive empathy processes), is an automatic process wherein the Other is given as experiencing, and often as experiencing a particular state. That is, rather than taking an object of perception and imagining or inferring my way to what it might be experiencing, I actually "directly perceive" (Zahavi, 2014; Zahavi and Rochat, 2015) its experience. For example, upon seeing a man crying, I might simply "perceive" that he is sad, without engaging in more conscious ("complex") empathy processes.

From a phenomenological perspective, empathy is fundamentally "how we experience others" (Zahavi, 2014, p. 130); it is the act "in which foreign experience is comprehended" (Stein, 1964, p. 6). Furthermore, this 'basic' class of empathy is a necessary precursor to, or component of, simulation-based and theorization-based complex empathy processes. In both instances—whether I am imagining or inferring the state of a target entity—I must first "grasp" the entity as an Other that is capable of experience. Indeed, I cannot be said to empathize with an entity unless I have already engaged with it as an experiencing Other, for—as the phenomenological perspective illustrates—empathy is precisely the experiencing of foreign experience. This "basic" or phenomenological class of empathy not only underpins the more "complex" forms, but as will be argued in **section 2.2**, it is also a necessary component of sympathy. In **section 3** of this paper I will argue that the basic kind of empathic engagement described here is an ethical engagement, such that in empathizing with an entity,[1] we have already engaged with it as possessing "minimal" moral status. Of the three forms of empathy discussed here, the phenomenological account has received the least attention in HRI.[2] However, HRI research on robot emotion expression (Kühnlenz et al., 2013; McColl and Nejat, 2014) could be understood as falling under the umbrella of basic empathy, in that the researchers aim to prompt users to perceive robots as experiencing certain states.

2.2 Defining Sympathy

As indicated in the **section 2.1**, sympathy is generally understood as feeling bad "for" an Other. Because of this, it is often conflated with pity—a term that has in recent history acquired a negative connotation (Nussbaum, 1996). "Pity" and "compassion," though not directly discussed in this paper, are closely related to sympathy—pity is best understood as a reduced form of sympathy, while "compassion" can be understood as describing a particularly strong instance of sympathy (Quick, 2021). However, upon closer examination, sympathy is a complex phenomenon that is closely related to compassion (Nussbaum, 2001) and is subject to complex social and interactional norms (Clark, 1997). Thus, I offer the following definition of sympathy:[3] (Quick, 2021):

[1] Note that because both complex empathy processes and sympathy are built upon this basic empathy process, the ethical engagement carries over into such interactions.

[2] This is not to say that phenomenological accounts of empathy have received *no* attention in HRI, for instance (Coeckelbergh, 2018) has also engaged with empathy, phenomenology, and robot moral status. As indicated in the introduction, I believe the approaches are compatible. Indeed, the conclusions reached by Coeckelbergh (and Gunkel, for that matter) are highly similar to those offered here, although the means of reaching these conclusions is different. The three approaches all emphasize that the phenomenology of human-robot interactions should be taken seriously. This account contributes to the discussion primarily in terms of an analysis of empathy and sympathy that supports the importance of the phenomenology of human-robot interactions and a reframing of the discussion in terms of implications for the design of empathetic and sympathetic robots.

[3] This definition of sympathy is drawn from (Quick, 2021).

Sympathy is a prosocial response **R** to the negative situation of an Other, which leads to an altruistic motivation, and whose appropriate expressions and instantiations are context-dependent and governed by dynamic social norms. **R** consists of several components, the first five of which are necessary, while the sixth is facultative:

i) Sentiment 'for'
ii) Some level of empathizing
iii) A judgment of seriousness
iv) A non-fault judgment
v) A value judgment

In addition, **R** may include:

vi) A specific behavioral display

In what follows, I will briefly[4] argue for the necessity of each of the components, beginning with the claim that sympathy is "subject to complex social and interactional norms." Candace Clark's extensive sociological research of American sympathy norms (Clark, 1987; Clark, 1997) suggests that sympathy is best understood in terms of *exchanges*—giving sympathy places an obligation of repayment on the Other, just as accepting sympathy places an obligation of repayment on oneself. Furthermore, sympathy exchanges need not occur in a one-to-one, universalizable fashion, they are instead always situated within a specific social context that dictates acceptable forms of displaying and repaying sympathy, as well as which sorts of circumstances merit sympathy. Sympathy "costs the donor time, effort, and emotional energy" (Clark, 1997, p. 130), and is thus a valuable commodity in our socio-emotional economy. Displays of sympathy are, as Arie Hochschild's work on emotions suggests, a form of "emotional labor" that is governed by "display rules" (Hochschild, 1983, p. 60). To be an effective sympathizer, one must understand—and comply with—the local sympathy norms.[5] A social agent that fails to act in accordance with these norms may be seen as what Clark has called a "sympathy deviant" (Clark, 1997, p. 22), eventually resulting in exclusion from the sympathy network.

The similarities between sympathy and compassion can be found in components (iii-v), which are drawn (and modified) from Martha Nussbaum's Aristotelian account of compassion, as well as Daniel Batson's social-psychological account of compassion (Batson, 2011). The judgment of seriousness indicates that for *genuine* sympathy to occur, the sympathizer must judge that the suffering of her target must be non-trivial or significant in some fashion. For instance, the suffering incurred by a paper cut is typically not seen as worthy of sympathy as the suffering incurred by losing a loved one. The non-fault judgment indicates that a sympathizer must judge that the victim is not responsible for his plight, or that if he is responsible, some extenuating circumstances mitigate this responsibility. Suppose I break my hand punching in a car window—without further information, my plight merits only minimal sympathy, if any. However, if we add that I punched in the window to rescue a baby who had been left in the hot car with closed windows for several hours, despite my being responsible for my injury, the altruistic intention behind the act can mitigate the importance of the fault, such that an observer may be more inclined to sympathize with me.

Finally, the value judgment ("eudaimonistic" judgment, in Nussbaum's terms (Nussbaum, 2001)), indicates that the object of one's sympathy must be seen as relevant to one's own flourishing—it must be "a significant element of my scheme of goals and projects, an end whose good is to be promoted" (Nussbaum, 2001, p. 321). Alternatively, in Batson's phrasing, I must "care about whether the other is in need and about how this need affects the other's life" (Batson, 2011, p. 41). Thus, in genuinely sympathizing, I will have judged (perhaps implicitly) that the suffering of the Other is serious, not of his or her own making (or justifiably so), and that the Other matters to me in some fashion. This "mattering" can take various forms. For instance, I need not explicitly judge that the Other—say, a robot who is being mistreated—is actually suffering, or indeed actually capable of suffering, but only that the robot appears to be suffering, while holding as a part of my 'scheme of goals and projects' a belief along the lines that "suffering is bad." As such, any entity which is perceived as suffering could be seen as relevant for my flourishing and judged as having value (at least initially—that is to say, judgments are subject to revision). It is here that perceptual, or phenomenological, empathy plays a particularly important role, in that it accounts for how we can perceive an entity (a robot, human, animal, etc.,) as suffering. For this reason, I argue that empathy "in some form" 2) is also a necessary component of sympathy—one cannot genuinely sympathize without first perceiving (or judging via inference or simulation) that the entity in question is suffering in some sense.[6] Likewise, it is constitutive of sympathy that one "feels for" 1) the victim. If I do not on some level feel (e.g., bad, or sad) for the victim, I cannot be said to genuinely sympathize.[7]

The sixth, facultative component of sympathy (display) is likely the most important in terms of human-robot sympathy exchanges, in that it seems to be, currently, the easiest and most impactful of these components to equip social robots with.

[4] The argumentation for this account of sympathy is per force brief, as the focus of this paper is on the implications sympathy and empathy have for robot moral status. For an extended discussion of various notions of sympathy see (Quick, 2021).

[5] For a further discussion of the norms and how they affect the design of sympathetic social robots, see (Quick, 2020).

[6] While Nussbaum has argued that empathy is not necessary for compassion, this seems to be because she limits the type of empathy considered to one that functions via simulation or imagination (Nussbaum, 2011, p. 149).

[7] One can also feel what might be called "routinized" sympathy, where one might have felt sympathy for an entity in the past but due to repeated expose no longer holds the judgments or sentiment in an "active" sense. For instance, I might actively sympathize with a homeless man the first time I see him on my way to my office, but over time come only to feel this "routinized" sympathy—active sympathy requires, as Clark indicates, "time, effort, and emotional energy" (Clark, 1997, p. 130).

Sympathy displays can take both what might be called *overt* and *subtle* forms. Overt displays of sympathy include acts such as verbal affirmations of sympathy ("I am sorry to hear that"), the giving of gifts such as flowers or money, or acts such as attending a funeral with a friend who has lost a loved one. Subtle displays, on the other hand, encompass acts such as sympathetic facial expressions or physical contact (e.g., placing a hand on a victim's shoulder). However, because of the complex and socially relative norms that govern when and how to express sympathy, sympathy displays can be seen as constituting the adoption of a social, political, or moral stance (i.e., by showing that one believes this particular plight is indeed worthy of sympathy). For instance, expressing sympathy for a woman who is unable to receive an abortion in Texas due to the 2021 anti-abortion legislation can be interpreted as adopting a "pro-choice" stance. Thus, the situations which merit sympathy displays, and the manners in which social robots ought to display sympathy—particularly in cross-cultural contexts—will need to be carefully considered (Quick, 2020).

Already in light of the preceding brief discussion of empathy and sympathy, it has become clear, I hope, why it is urgent that the concepts are employed with further care than is often seen in HRI. An experiment that measures empathy in cognitive terms is not immediately comparable to an experiment which measures empathy in affective terms, nor are they comparable to an experiment that purports to measure empathy but actually measures sympathy. In addition, more careful attention to the cognitive and emotional processes involved in these two different phenomena, empathy and sympathy, can prove decisive for the discussion of whether robots can or should have moral standing, and ensuing recommendations for the (physical-kinematic and functional) design of the robot. While the problem of mixing empathy and sympathy is not unique to HRI research, the increasing interest in commercial and domestic social robots lends an urgency to the task of understanding the social and moral implications of social robots that elicit or display sympathy and empathy that is simply not found in many areas of research. For instance, whether philosophers and psychologists agree on the nature of empathy and sympathy in 10 years or in one hundred years makes relatively little difference in practical terms. Such research may indeed result in social benefits (e.g., improved techniques in therapy or pedagogy), but a failure to reach conclusions here, or a delay in doing so, will at least not actively cause harm. The same cannot be assumed in the case of sympathetic and empathetic robots. Because the development of such devices is still in early stages, it is not clear what ethical, social, or emotional impact such devices may have on their users.

3 ROBOT MORAL STANDING

I will argue in this section that empathizing—in any of the three senses discussed—involves an ascription of what we might consider "minimal" moral standing, while sympathizing involves a still greater ascription of moral standing. The argument for why empathizing with a robot entails an ascription of moral status can be seen as proceeding from four premises (Quick, 2021), each of which I will argue for by drawing primarily on the works of the phenomenologists Edith Stein (1964) and Max Scheler (1923). These are:

1. The feelings a human may have for, or on behalf of, a robot are genuine experiences of the same kind as those a human may have for, or on behalf of, another human, regardless of the robot's (lack of) internal states.
2. The human experience of foreign experience or, more precisely, of "an experiencing Other," is[8] of one kind, regardless of the ontological status of the Other.
3. Actions perceived as intentional are apprehended as originating from an experiencing Other.
4. Others apprehended as experiencing are due moral consideration.

3.1 Experiencing Otherness and Moral Standing

The first premise can be traced to Edith Stein's account of empathy, wherein she argued that while we may be deceived with regards to the object of our feelings, we cannot be deceived as to the existence of the feelings themselves. "I can be deceived in the object of my love, i.e., the person I thought I comprehended in this act may in fact be different, so that I comprehended a phantom. But the love was still genuine" (Stein, 1964, p. 31). In other words, even though the object of my love was not reciprocating the feeling, was unfeeling with respect to love or not as I initially comprehended it to be, my own feeling of love towards the Other was still genuine. With this, we can understand the results of HRI research such as Bartneck and Hu's 2008 Milgram experiments, wherein the researchers noted that "the participants showed compassion for the robot" (Bartneck and Hu, 2008, p. 420).[9] The sympathy (or compassion) that participants felt towards the robotic victim was of the same kind that participants in the original Milgram experiment might have felt for the human victims, and just as genuine, regardless of the fact that the robots were not actually suffering. Thus, even if a participant came to know that the robot was not actually suffering (and was in fact incapable of suffering), the empathic experience he or she had of the Other as "experiencing pain" remains genuine.

This leads directly to the second premise, which argues that our experience of an entity as an Other that is experiencing (or is capable of experiencing) mental or affective states is not tied to

[8] The usage of "experiencing Other" rather than simply "Other" is intended to reflect that there may be other possible forms of "otherness," such as "logical otherness," which are not given through empathy. The otherness given through empathy will always be 'experiencing otherness', for, as indicated in **section 2.1**, empathy is simply the comprehension of foreign experience (Stein, [1919/1964] 1989, p. 6).

[9] Note that while participants in this study may have shown "compassion," they still followed through and applied the maximum voltage. This does not, however, indicate that participants did not genuinely sympathize, only that they did not overtly display sympathy by refusing to continue.

the actual or perceived ontological status of that entity. In other words, whether a robot is actually capable of suffering or not—or whether the observer believes it to be capable of suffering or not—it is entirely possible for one to experience the robot as suffering. The insights of early phenomenologists that "experiencing X as suffering" is independent of the ontological state of X also seems to underlie Mark Coeckelbergh's argument that "whatever the 'real' status of the robot may be, it is its appearance that is relevant to how the human-robot relation is experienced and constructed" (Coeckelbergh, 2011, p. 198). Unsurprisingly, given his use of the phenomenology of Emmanuel Levinas, a similar thread can be found in David Gunkel's work, when he argues that rather than first identifying the ontological status of an entity and then deciding on its moral status, "we are first confronted with a mess of anonymous others who intrude on us and to whom we are obligated to respond even before we know anything at all about them" (Gunkel, 2018b, p. 96). In sum, the first and second premises can be taken as suggesting that the human empathic experience of otherness (i.e., the experience that a particular Other is capable of experience or is experiencing an affective or mental state) is not contingent on ontological knowledge and is, as such, not to be understood as a perceptual mistake—as experience, it is a correct processing of the data (Quick, 2021, p. 258).

Of course, some objects and entities might lend them themselves to being experienced empathically as Other more readily than others. Two possible reasons for this are as follows. First, it could be that there are certain affordances or characteristics that we recognize in objects as being associated with Otherness. For instance, Stein discusses what she calls "the specific phenomena of life," which include "growth, development and aging, health and sickness, vigor and sluggishness" (Stein, 1964, p. 68). As she indicates, it is not merely that we attach these characteristics to an object after perceiving it, but rather, that through the act of empathy they are "co-seen"— "Thus, by his walk, posture, and his every movement, we also "see" "how he feels," his vigor, sluggishness, etc." (Stein, 1964, p. 69). Certain objects, such as humans, animals, and social robots, simply present themselves as experiencing these states more clearly than objects such as rocks or guitars. Additionally, another key difference between objects such as rocks and robots is simply the fact that social robots (often) possess some movement capabilities. More precisely, they can present themselves as capable of voluntary movement in a way that rocks simply cannot.[10] In line with this, a second reason for why we empathize more readily with some objects than others could have to do with similarity to previous Others. That is, if an object is similar to, shares sufficient characteristics with, or in some meaningful way reminds me of one that I have previously grasped as Other, I may be predisposed to grasp it as such than if it did not. For example, a humanoid robot may be more readily grasped as Other simply because it bears a resemblance to the "standard" Other—humans. A rock, on the other hand, does not bear much of a resemblance to humans, or animals, or social robots—thus, it may be less predisposed to grasping it as Other through empathy.

The third premise holds that actions which are perceived as intentional are perceived (perhaps implicitly) as originating from an experiencing Other. That is, if we understand an action as intentional, then we are understanding it as an action that is underpinned by a volition, intention, or willing.[11] While the nature of these three concepts is debatable, they are all undoubtably experiential in some sense, such that an agent which is incapable of experiencing is incapable of willing or having intentions or volitions in the way that humans are. Despite believing this, we often engage with agents—such as social robots—*as if* they are acting intentionally,[12] or *as if* they are experiencing. Regardless of whether (or not) participants explicitly believe a social robot possesses mental states, intentions, or experiences, humans often seem to engage with them as if they do, going so far as to feel bad for them when they are "suffering" (Bartneck and Hu, 2008; Darling et al., 2015; Seo et al., 2015; Carlson et al., 2019;). If it is the case that intending and willing are a form of experiencing, then we can see that robotic actions which are perceived as intentional are perceived as originating from an experiencing Other—for, as Scheler wrote, ". . .we cannot be aware of an experience without being aware of a self. . ." (Scheler and Health, 1923, p. 9). With respect to the current discussion, we could modify this to say that "we cannot perceive an experience without perceiving an Other" Similarly, Stein argued that "willing is essentially motivated by a feeling" (Stein., 1964, p. 97) and "the foreign person is constituted in empathically experienced acts. I experience his every action as proceeding from a will and this, in turn, from a feeling" (ibid. 109).

The fourth premise is a normative claim that any entity which is apprehended as an experiencing Other is due some level of moral consideration. It is with regards to this claim in particular that the phenomenological account of empathy offers something novel to the current debate about robot moral standing, which has largely centered around the Kantian, utilitarian, and virtue ethics based answers to the problem.[13] The phenomenological perspective offers an epistemological argument—in not opening ourselves to the full datum (experiencing the Other as experiencing *and* worthy of moral consideration) we are making an experiential mistake. That is, a robot simulating experiential states is "correctly" experienced when it is experienced as an experiencing Other, and qua this, also as due moral

[10] Stein indicates that while voluntary movement is a key aspect of ascribing Otherness to an entity, it is not strictly necessary; for instance, we can empathize, in a limited sense with plants, and recognize them as "alive," without ascribing them consciousness or states such as pain and pleasure (Stein, 1964, p. 69).

[11] For a discussion of intentions and volitions, see (Adams and Mele, 1992). The focus here will lie on intentions and the will—as Adams and Mele argue, "volition" does not seem to add much to the 'intention' concept.

[12] See (Seibt, 2017) for a discussion of "as if," and (Dennett, 1995) for a discussion of humans engaging with objects in this manner.

[13] Cf. (Coeckelbergh, 2011; Gunkel, 2012).

consideration. The argument can be framed in terms of Scheler's discussion of brutality, which is understood as the "disregard of other peoples' experience, despite the apprehension of it in feeling" (Scheler and Health, 1923, p. 14). Furthermore, "to regard a human being as a mere log of wood and to treat the object accordingly is not to be "brutal" towards him" (ibid.)—we are only brutal in cases where we apprehend an entity as an experiencing Other yet do not extend moral consideration to it. If an object is genuinely seen as an unintentional, non-experiencing object and treated as such, then it seems we are not engaging in brutality. Likewise, when a robot is genuinely experienced as non-experiencing, as non-Other, not including it in our moral considerations is not a moral failure, it is a correct processing of the data. The situation under consideration here is one where the robot or entity *is* experienced, through empathy, as an Other—it is here that moral consideration is due.[14] However, this brings us to a second, closely related and similarly morally objectionable act in which we deny the experience of an entity; namely, dehumanization. In dehumanizing a person, we may ascribe fewer human attributes to them, or go so far as to ascribe "deficient or absent humanity to a target" (Haslam and Loughnan, 2014, p. 406). In regarding a human as a "mere log of wood," we are dehumanizing her, stripping away her experiencing otherness—we are not brutalizing her, for we did not first empathically experience her as an experiencing Other.[15]

Dehumanizing is clearly morally objectionable for various reasons, but I will argue that one of the factors that makes it "wrong" is related to that which makes brutality wrong—the denial (or disregard) of experiencing otherness. To further investigate this claim, we can adopt a distinction between two types of capacities: "agency" (i.e., cognitive capacities such as planning and thought) and 'experience' (i.e., capacities such as emotions and consciousness) (Gray et al., 2007). While Gray et al. used "experience" to indicate a specific set of capacities which are distinct from those that fall under the "agency" category, per the arguments discussed in relation to the third premise, it appears that agentic capacities are experiential. For instance, take the standard understanding of "thinking"—this can be understood as being an intentional act, or as motivated by the will or feelings directly, or as a phenomenal act, in that there is there is "something that it is like" to "think." When we say a non-human object, such as a robot, is "thinking," we are either simply using figurative language, or anthropomorphizing the robot. In the latter case, we are perceiving the robot as an experiencing Other and, according to the preceding arguments, we have incurred an obligation to extend at least some moral consideration to the robot. Such a position is compatible with, and provides further support for, virtue-based arguments for extending moral consideration to robots. On such a view, "mistreating a robot is not wrong because of the robot, but because doing so repeatedly and habitually shapes ones moral character in the wrong kind of way... Mistreating the robot is a vice" (Coeckelbergh, 2018, p. 145).

As suggested earlier, the experience of foreign experience (i.e., empathy, in the phenomenological sense) is not merely a perceptual mistake; rather, it is a way of being "true to the situation." Incorporating a phenomenological perspective of empathy thus introduces a methodological switch for the discussion of robot moral standing. Instead of considering the acts of the subject in relation to a preconceived ontology of the object, and thereby sorting our perceptions as "accurate" or "inaccurate," the phenomenologist analyzes "what is given in experience." On such a view, we can see that what occurs in brutality and dehumanization is a failure to take in fully that which is "there for experiencing." Social robots that simulate experiential states can create the same sort of experiential data as humans do, and a rejection of this data is an experiential (and ethical) error of the same sort that brutalizing or dehumanizing a human would be—it is a rejection of the "phenomenological truth" which confronts me.

3.2 A Return to Sympathy

From the discussion of sympathy and empathy found in the previous sections, it seems that sympathy is of greater importance for the debate of robot moral standing than empathy is (particularly in terms of cognitive and affective empathy). For one thing, sympathy includes empathy as a necessary component, such that if I sympathize with a robot, I have already empathized with it. At this point, I have already framed the robot as an experiencing Other and ascribed a "minimal" moral status, in that I have incurred an obligation to take the experiential data given through my empathy seriously. Sympathizing with a robot, however, requires that I engage with it as an experiencing Other to an even greater extent. I must consider whether the robot's "suffering" is of its own making, whether it is serious, and perhaps most importantly for questions of moral standing, I must judge the robot (or it's suffering) as important and relevant to my own flourishing in some sense. It is unsurprising then that the focus of current HRI research on sympathy in human-robot interactions has typically been on whether humans can have sympathy for robots. While this is certainly an important question, the discussion does not move the robot beyond the status of a potential moral patient. An investigation of situations

[14] One might be concerned with a situation in which we experience the robot as Other, but also know that it is *not* an Other (i.e., non-experiencing). Indeed, such a case appears similar to when I have a fear that I know to be irrational, for instance, when I experience fear with respect to the monster in my closet, despite knowing there is no monster in the closet. Such fear is genuine as an experience but ought not dictate our actions, given its irrationality. However, the cases are not actually the same—in the case of fearing the monster while knowing it does not exist or that the fear is unfounded, not letting the fear dictate my actions leans more towards being a virtuous act than a vicious one. In overcoming that fear I practice the process of being courageous, whereas in 'overcoming' my perception of the robot as an experiencing Other I am practicing a vicious process, namely dehumanization, as it is discussed in the remainder of this section. Thanks to an anonymous reviewer for raising this concern.

[15] Indeed, dehumanization seems to be a failure to recognize the Other which is given to us *as* an Other. In this sense, it is relatable to the preceding discussion of why we may be more inclined to empathize with some objects than others. If our previous classifications of entities as Other/non-Other influences our future classifications as such, then it is sensible that entities such as social robots, which do not fit neatly into our existing categories may be experienced as Other or non-Other with a greater degree of variation than objects that have more stable categorizations as non-Other (e.g., rocks).

in which a robot shows sympathy for a human, on the other hand, would move us into the realm of considering whether robots can be potential moral agents.

A robot which displays sympathy for a human is potentially a moral agent in that it presents itself as an entity that is capable of experiencing foreign experience, as well as one which understands the local sympathy norms (at least in so far as it is able to comply with them). Despite the importance of investigating sympathy displaying robots, the topic has received little attention,[16] perhaps in part due simply to technological limitations. A robot which is able to display sympathy convincingly or meaningfully will require reliable affect recognition as well as a set of rules for the sorts of situations that require displays and a library of potential displays that are linked to specific classes of situations. Indeed, as Kerstin Fischer suggests, "When we speak of robots processing and using social signals, then we are discussing future technologies" (Fischer, 2019, p. 19). The investigation of what sort of status is due to a robot which displays sympathy raises a variety of questions for future research. For instance, Clark's research suggests that sympathy requires reciprocity, such that we can predict that when we sympathize with a robot (in the context of a long-term interaction), we will eventually expect 'repayment'. In human-human relations, one of the principal forms of repaying the sympathy someone offers you is with an offer of sympathy (at a future, appropriate, time). Thus, an effective sympathetic robot, for instance, one that is intended to act as a "companion," will require the ability to offer sympathy as well as accept it if it is to function as an effective actor within our sympathy networks (Quick, 2020). Indeed, if Clark's observations regarding the expectation of reciprocity in sympathetic interactions between humans holds true in the case of robots, we should investigate what sort of threat a robot which elicits sympathy—without sympathizing in turn—poses to our sympathy conventions.

As argued in the previous section, when we sympathize with a robot, we are not making a "sentimental mistake" rather, we are avoiding precisely such an experiential mistake (brutalization or dehumanization) by being open to the available phenomenological truth. However, when it comes to a robot's display of sympathy, we must ask whether this "truth" is present in the same manner—are we experiencing a foreign experience of foreign experience in the way that we can with a human's display of sympathy? That is, can a robot's apparent sympathy for a human be empathically experienced as genuine—in the way that data from HRI has suggested that a robot's suffering can—or will it always be perceived as a simulation of sympathy? In sum, the ethical debate about the moral standing of robots appears to be miscalibrated. The focus should not be on whether Kantian or virtue-ethical arguments are better for justifying "attributions" of moral standing, but should rather be on: how much do we want to threaten our sympathy conventions? Our empathic engagement with the robot already indicates an ethical engagement with it, in that we have experienced it as an experiencing Other. Is it preferable to have social robots that we can genuinely sympathize with—to open ourselves to what is given in experience, the datum of foreign experience—but which will not show sympathy? Or should robots which elicit sympathy also show sympathy, even though it may be perceived as inauthentic? These questions are very different than those which are typically discussed in relation to robot moral standing and are of a more empirical than normative nature.

4 CONCLUSION

In this paper, I have argued that the debate over empathy in human-robot interactions has largely failed to recognize the distinctions between the three types of empathy on the one hand, and sympathy on the other. Furthermore, the phenomenological account of empathy, which offers critical insights and valuable research avenues into the question of robot moral standing, has largely been overlooked. This type of empathy, namely empathy as the "experience of foreign experience" is not only central to other forms of empathy (such as affective and cognitive empathy) as well as sympathy, but also explains best the connection between empathy, sympathy, and moral standing. Additionally, I argued for a novel account of sympathy which attempts to clarify the distinction between empathy and sympathy and outline the necessary conditions for an instance of genuine sympathy. In relation to this, two types of experiential errors—brutality and dehumanization—were discussed, and it was argued that both represent a failure to properly consider the data provided through our empathetic and sympathetic experiences. While the phenomenological analysis of empathy and the account of sympathy that have been discussed here offer a way of reframing the question of robot moral status, they also lead to a wide range of new questions for HRI research, several of which were posed in section three. Recognizing that our empathic engagement with an Other already also constitutes an ethical engagement with it allows for us to move from the heavily discussed normative questions to novel ones, as well as conduct empirical research on to what extent humans feel and respect the moral obligations which result from engaging with social robots that display (and elicit) differing levels of affect and sympathy.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

Research resulting this paper was funded by the Carlsberg Foundation (Grant nr. CF16-0004).

[16] There are, of course, exceptions (e.g. (Leite et al., 2014)).

REFERENCES

Adams, F. (2001). Empathy, Neural Imaging and the Theory versus Simulation Debate. *Mind Lang.* 16 (4), 368–392. doi:10.1111/1468-0017.00176

Adams, F., and Mele, A. R. (1992). The Intention/Volition Debate. *Can. J. Philos.* 22 (3), 323–337. doi:10.1080/00455091.1992.10717283

Asada, M. (2015). Development of Artificial Empathy. *Neurosci. Res.* 90, 41–50. doi:10.1016/j.neures.2014.12.002

Baron-Cohen, S., and Wheelwright, S. (2004). The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *J. Autism Dev. Disord.* 34 (2), 163–175. doi:10.1023/B:JADD.0000022607.19833.00

Bartneck, C., and Hu, J. (2008). Exploring the Abuse of Robots. *Is* 9 (3), 415–433. doi:10.1075/is.9.3.04bar

Batson, C. D. (2011). *Altruism in Humans*. 1st ed.. Oxford, UK: Oxford University Press.

Bloom, P. (2018). *Against Empathy: The Case for Rational Compassion*. Reprint edition. NY, USA: Ecco.

Carlson, Z., Lemmon, L., Higgins, M., Frank, D., Salek Shahrezaie, R., and Feil-Seifer, D. (2019). Perceived Mistreatment and Emotional Capability Following Aggressive Treatment of Robots and Computers. *Int. J. Soc. Robotics* 11 (5), 727–739. doi:10.1007/s12369-019-00599-8

Carruthers, P., and Smith, P. (1996). *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Ceh, S., and Vanman, E. J. (2018). The Robots Are Coming! the Robots Are Coming! Fear and Empathy for Human-like Entities. *PsyArXiv*. doi:10.31234/osf.io/4cr2u

Clark, C. (1997). *Misery and Company: Sympathy in Everyday Life*. Chicago, US: University of Chicago Press.

Clark, C. (1987). Sympathy Biography and Sympathy Margin. *Am. J. Sociol.* 93 (2), 290–321. doi:10.1086/228746

Coeckelbergh, M. (2012). *Growing Moral Relations: Critque of Moral Status Ascriptions*. London, UK: Palgrave Macmillan.

Coeckelbergh, M. (2011). Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. *Int. J. Soc. Robotics* 3 (2), 197–204. doi:10.1007/s12369-010-0075-6

Coeckelbergh, M. (2018). Why Care about Robots? Empathy, Moral Standing, and the Language of Suffering. *Kairos. J. Philos. Sci.* 20 (1), 141–158. doi:10.2478/kjps-2018-0007

Darling, K., Nandy, P., and Breazeal, C. (2015). "Empathic Concern and the Effect of Stories in Human-Robot Interaction," in 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Kobe, Japan, August 31-September 4, 2015, 770–775. doi:10.1109/ROMAN.2015.7333675

de Vignemont, F., and Singer, T. (2006). The Empathic Brain: How, when and Why? *Trends Cogn. Sci.* 10 (10), 435–441. doi:10.1016/j.tics.2006.08.008

Dennett, D. C. (1995). *The Intentional Stance*. Cambridge, UK: MIT Press.

Fischer, K. (2019). Why Collaborative Robots Must Be Social (And Even Emotional) Actors. *Techné: Res. Philos. Tech.* 23 (3), 270–289. doi:10.5840/techne20191120104

Goldman, A. I. (2006). "Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading," in *Simulating Minds* (Oxford, UK: Oxford University Press).

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of Mind Perception. *Science* 315 (5812), 619. doi:10.1126/science.1134475

Gunkel, D. J. (2018a). *Robot Rights*. Cambridge, UK: MIT Press.

Gunkel, D. J. (2012). *The Machine Question*. Cambridge, UK: The MIT Press.

Gunkel, D. J. (2018b). The Other Question: Can and Should Robots Have Rights? *Ethics Inf. Technol.* 20 (2), 87–99. doi:10.1007/s10676-017-9442-4

Haslam, N., and Loughnan, S. (2014). Dehumanization and Infrahumanization. *Annu. Rev. Psychol.* 65 (1), 399–423. doi:10.1146/annurev-psych-010213-115045

Hochschild, A. R. (1983). *The Managed Heart: Commercialization of Human Feeling (Updated, with a New Preface)*. Berkeley, California, US: University of California Press.

Hollan, D. (2012). Emerging Issues in the Cross-Cultural Study of Empathy. *Emot. Rev.* 4 (1), 70–78. doi:10.1177/1754073911421376

Hume, D. (1740). in *A Treatise of Human Nature*. Editors D. Norton and M. Norton (Oxford, UK: Oxford University Press), 1.

Kühnlenz, B., Sosnowski, S., Buß, M., Wollherr, D., Kühnlenz, K., and Buss, M. (2013). Increasing Helpfulness towards a Robot by Emotional Adaption to the User. *Int. J. Soc. Robotics* 5 (4), 457–476. doi:10.1007/s12369-013-0182-2

Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2014). Empathic Robots for Long-Term Interaction. *Int. J. Soc. Robotics* 6 (3), 329–341. doi:10.1007/s12369-014-0227-1

Leite, I. (2015). Long-term Interactions with Empathic Social Robots. *AI Matters* 1 (3), 13–15. doi:10.1145/2735392.2735397

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., and Paiva, A. (2013). The Influence of Empathy in Human-Robot Relations. *Int. J. Human-Computer Stud.* 71 (3), 250–260. doi:10.1016/j.ijhcs.2012.09.005

McColl, D., and Nejat, G. (2014). Recognizing Emotional Body Language Displayed by a Human-like Social Robot. *Int. J. Soc. Robotics* 6 (2), 261–280. doi:10.1007/s12369-013-0226-7

Menne, I. M., and Schwab, F. (2018). Faces of Emotion: Investigating Emotional Facial Expressions towards a Robot. *Int. J. Soc. Robotics* 10 (2), 199–209. doi:10.1007/s12369-017-0447-2

Nussbaum, M. C. (2011). "Compassion: Human and Animal," in *Species Matters: Humane Advocacy and Cultural Theory* (NY, USA: Columbia University Press), 240. doi:10.7312/deko15282-007

Nussbaum, M. (1996). Compassion: The Basic Social Emotion. *Soc. Phil Pol.* 13 (1), 27–58. doi:10.1017/S0265052500001515

Nussbaum, M. C. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511840715

Preston, S. D., and de Waal, F. B. M. (2002). Empathy: Its Ultimate and Proximate Bases. *Behav. Brain Sci.* 25 (1), 1–20. doi:10.1017/S0140525X02000018

Quick, O. S. (2020). Challenges for Sympathetic Robot Design. *Culturally Sust. Soc. Robotics: Proc. Robophilosophy* 2020, 335. doi:10.3233/faia200929

Quick, O. S. (2021). *Sympathizing and Empathizing with the Robotic Other*. Ph.D. thesis. Aarhus(Denmark): Aarhus University.

Ravenscroft, I. (1998). What Is it like to Be Someone Else? Simulation and Empathy. *Ratio* 11 (2), 170–185. doi:10.1111/1467-9329.00062

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., and Robinson, P. (2009). "Empathizing with Robots: Fellow Feeling along the Anthropomorphic Spectrum," in 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, Netherlands, September 10-12, 2009, 1–6. doi:10.1109/ACII.2009.5349423

Rosenthal-von der Pütten, A. M., Schulte, F. P., Eimler, S. C., Sobieraj, S., Hoffmann, L., Maderwald, S., et al. (2014). Investigations on Empathy towards Humans and Robots Using fMRI. *Comput. Hum. Behav.* 33, 201–212. doi:10.1016/j.chb.2014.01.004

Scheler, M., and Heath, P. (19232008). *The Nature of Sympathy (Wesen und Formen der Sympathie)*. 1st ed.. Piscataway, US: Transaction Publishers.

Seibt, J. (2017). "Towards an Ontology of Simulated Social Interaction: Varieties of the "As if" for Robots and Humans," in *Sociality and Normativity for Robots*. Editors R. Hakli and J. Seibt (NY, USA: Springer International Publishing), 11–39. doi:10.1007/978-3-319-53133-5_2

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., and Young, J. E. (2015). "Poor Thing! Would You Feel Sorry for a Simulated Robot?: A Comparison of Empathy toward a Physical and a Simulated Robot," in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Portland, Oregon, USA, March 2-5, 2015, 125–132. doi:10.1145/2696454.2696471

Singer, T. (2006). The Neuronal Basis and Ontogeny of Empathy and Mind reading: Review of Literature and Implications for Future Research. *Neurosci. Biobehavioral Rev.* 30 (6), 855–863. doi:10.1016/j.neubiorev.2006.06.011

Smith, A. (1759). *The Theory of Moral Sentiments (Kindle Ebook)*. Boston¸US: Digireads.com Publishing.

Stein, E. (1964). *On the Problem of Empathy (3. Rev. ed., Vol. 3)*. Washington, DC, US: ICS Publications.

Stephan, A. (2015). Empathy for Artificial Agents. *Int. J. Soc. Robotics* 7 (1), 111–116. doi:10.1007/s12369-014-0260-0

Stueber, K. (2019). "Empathy," in *The Stanford Encyclopedia of Philosophy (Fall 2019)*. Editor E. N. Zalta (Stanford, USA: Metaphysics Research Lab, Stanford University).

Stueber, K. (2006). *Rediscovering Empathy Agency, Folk Psychology, and the Human Sciences*. 1st ed.. Cambridge, UK: MIT Press.

Zahavi, D., and Rochat, P. (2015). Empathy≠sharing: Perspectives from Phenomenology and Developmental Psychology. *Conscious. Cogn.* 36, 543–553. doi:10.1016/j.concog.2015.05.008

Zahavi, D. (2014). *Self and Other: Exploring Subjectivity, Empathy, and Shame*. 1st ed.. Oxford, UK: Oxford University Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 04 January 2022
doi: 10.3389/frobt.2021.788242

Does the Correspondence Bias Apply to Social Robots?: Dispositional and Situational Attributions of Human Versus Robot Behavior

Autumn Edwards and Chad Edwards*

Communication and Social Robotics Labs, School of Communication, Western Michigan University, Kalamazoo, MI, United States

Increasingly, people interact with embodied machine communicators and are challenged to understand their natures and behaviors. The Fundamental Attribution Error (FAE, sometimes referred to as the correspondence bias) is the tendency for individuals to over-emphasize personality-based or dispositional explanations for other people's behavior while under-emphasizing situational explanations. This effect has been thoroughly examined with humans, but do people make the same causal inferences when interpreting the actions of a robot? As compared to people, social robots are less autonomous and agentic because their behavior is wholly determined by humans in the loop, programming, and design choices. Nonetheless, people do assign robots agency, intentionality, personality, and blame. Results of an experiment showed that participants made correspondent inferences when evaluating both human and robot speakers, attributing their behavior to underlying attitudes even when it was clearly coerced. However, they committed a stronger correspondence bias in the case of the robot–an effect driven by the greater dispositional culpability assigned to robots committing unpopular behavior–and they were more confident in their attitudinal judgments of robots than humans. Results demonstrated some differences in the global impressions of humans and robots based on behavior valence and choice. Judges formed more generous impressions of the robot agent when its unpopular behavior was coerced versus chosen; a tendency not displayed when forming impressions of the human agent. Implications of attributing robot behavior to disposition, or conflating robot actors with their actions, are addressed.

Keywords: fundamental attribution error, correspondence bias, social robot, human-robot interaction, computers are social actors, behaviorism

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Tom Ziemke,
Linköping University, Sweden
Aleksandra Swiderska,
University of Warsaw, Poland

****Correspondence:***
Autumn Edwards
autumn.edwards@wmich.edu

Specialty section:
This article was submitted to Ethics in Robotics and Artificial Intelligence,
a section of the journal Frontiers in Robotics and AI

Received: *01 October 2021*
Accepted: *22 November 2021*
Published: *04 January 2022*

Citation:
Edwards A and Edwards C (2022) Does the Correspondence Bias Apply to Social Robots?: Dispositional and Situational Attributions of Human Versus Robot Behavior. Front. Robot. AI 8:788242. doi: 10.3389/frobt.2021.788242

1 INTRODUCTION

The Fundamental Attribution Error (FAE) is the tendency for people to over-emphasize dispositional or personality-based explanations for others' behavior while under-emphasizing situational explanations (Ross, 1977). In other words, people sometimes demonstrate a cognitive bias by inferring that a person's actions depend on what "kind" of person they are rather than on the social and environmental forces that influence the person. As such, an observer will likely attribute reasons for a behavior to internal characteristics and not external factors (Gilbert and Jones, 1986).

Individual behavior is heavily influenced and guided by situational and external factors. However, "because people are accustomed to seeing individuals as causal agents, viewing the actor and (their) actions as forming a single categorical unit also appears to be the simplest, most satisfying, and least effortful inferential strategy (Heider and Simmel, 1944; Heider, 1958; Jones, 1979)" (Forgas, 1998, p. 319).

Although this effect has been thoroughly examined with humans, we do not know if the same correspondence bias will apply to social robots. When communicating with machines such as social robots, people must form impressions of the agents and judge their behavior. Compared to people, current robots are less agentic and autonomous with behaviors driven by programming, design, and humans in the loop. However, people do nonetheless assign robots agency, intentionality, and blame (Sciutti et al., 2013; De Graaf and Malle, 2019; Banks, 2020). The purpose of this experiment is to determine whether people commit the FAE in response to the behaviors of a social robot. FAE is sometimes referred to as the correspondence bias (Gawronski, 2004), an issue we will return to in the discussion. Whereas the FAE assumes a general tendency to underestimate the power of situation on human behavior, the correspondence bias refers more narrowly to the tendency to make disposition-congruent inferences of observed behavior. However, because much of the literature uses both FAE and correspondence bias, we will use the terminology cited in the mentioned studies in the next sections.

2 FUNDAMENTAL ATTRIBUTION ERROR

Research has demonstrated that the FAE may distort an observer's judgment of an individual, especially in the case of overattribution of individual responsibility for large achievements or grave mistakes (Ross et al., 1977). Previous research has demonstrated that individuals who commit the FAE assign too much personal responsibility for both positive and negative outcomes (Ross et al., 1977; Riggio and Garcia, 2009). According to research on the FAE, individuals use two types of information when making attributions: dispositional and situational (Pak et al., 2020). As such, the FAE "rests on an assumption of dualism: that there is a clear division between what is inside and outside the person" (Langdridge and Butt, 2004, p. 365).

Dispositional attributions pertain to perceived qualities of the individual, whereas situational attributions pertain to perceived characteristics of the environment and factors outside of the individual's control. "Potential biases in the causal attribution process can come from the valence of the situational outcome (was the outcome positive or negative), the degree of informational ambiguity of the situation, and the degree of control an actor has over an outcome" (Pak et al., 2020, p. 422). FAE has been examined in relation to behavioral judgments. For example, when presented with an excerpt of a character's bad day, students tended to attribute the cause to dispositional versus situational factors (Riggio and Garcia, 2009). However, students who were primed by watching a video about the power of social and environmental influences on individual behavior attributed the cause of the bad day more to situational factors. Therefore, broader construal may help attenuate the FAE.

FAE does not seem to be universal across cultures but does exist heavily in Western cultures (Norenzayan and Nisbett, 2000). Research in social psychology has forwarded several explanations for why individuals commit the FAE. The first explanation is that people are more likely to attribute causes or responsibilities to an observed than an unobserved element. Because agents are more salient than their situations in many judgment tasks, the agent itself draws observers' attributional focus (Taylor and Fiske, 1975; Robinson and McArthur, 1982). The second explanation is that personal/dispositional attributions are more comforting causal inferences because they reinforce the just-world hypothesis, which holds that "people get what they deserve" or "what goes around comes around" (Walster, 1966). However, this explanation better explains deliberative judgments than the swift or automatic judgments often formed in response to individual behavior (Berry and Frederickson, 2015).

The third explanation is that humans may have evolved (and learned) to be hypersensitive in terms of agency detection. HADD, or the hypersensitive agency detection device, is the cognitive system theorized to be responsible for detecting intentional agency (Barrett, 2000). People overestimate the presence of human agency and therefore demonstrate a bias in which situations and events are attributed to people or other human-like entities. Agency detectors are so sensitive that even movement is enough to trigger attributions of will and intention, as evidenced in a number of Theory of Mind (ToM) studies (Barrett, 2007).

2.1 Attributional Process in Human-Robot Interaction

People attribute mental states to others in order to understand and predict their behavior. There is evidence of similarity in how people interpret humans' and robots' actions in the sense that people implicitly process robots as goal-oriented agents (Sciutti et al., 2013), use the same "conceptual toolbox" to explain the behavior of human and robot agents (De Graaf and Malle, 2019), make implicit Theory of Mind (ToM) ascriptions for machine agents (Banks, 2020), and evaluate a social robot's message behavior in terms of its underlying beliefs, desires, and intentions for communication (Edwards et al., 2020). HRI scholars have argued that the physical presence of a robot, or embodied machine agent, can produce patterns of attributions similar to those occurring in human-human interaction (Ziemke et al., 2015; De Graaf and Malle, 2017; Pak et al., 2020). Even when participants were provided with transparent information about how a robot makes decisions, they still attributed outcomes of behaviors to robot thinking (Wortham et al., 2017), which suggests the persistence of dispositional attributions even when situational information is provided (Pak et al., 2020). In addition, people have been found to use folk-psychological theories similarly to judge human and robot behavior in terms of ascriptions of intentionality, controllability, and desirability and in the perceived plausibility of behavior explanations (Thellman et al., 2017). Furthermore, there is evidence that human-linked stereotype activation (e.g., stereotypes of aging) influences causal attributions of robot behavior (Pak et al., 2020). The results of such studies generally lend support to the Computers are Social Actors (CASA) paradigm, which posits that people tend to treat and respond to machine agents with social cues in the same ways they do other people (Reeves and Nass, 1996).

The Form Function Attribution Bias (FFAB) refers to cognitive shortcuts people take based on the robot's morphology or appearance (Haring et al., 2018). The FFAB leads people to make biased interpretations of a robot's ability and function based on the robot's physical form (Hegel et al., 2009) and the perceived age of the robot (Branyon and Pak, 2015). Some research has demonstrated that attributions of action and mind increased as more human features were included in pictures of robot/avatar faces (Martini et al., 2016). Interacting with robotic agents on a task reduced one's own sense of agency similar to working with other individuals (Ciardo et al., 2020). This effect was not observed with non-agentic mechanical devices. Other research suggests that agent-category cues help shape perceptions which then influence behavioral outcomes (Banks et al., 2021). In doing so, there is a tendency to judge action on the basis of the agent performing it. Although these findings do not speak directly to the applicability of the FAE to social robots, they do demonstrate that attributional patterns similar to those observed in human interaction may emerge when people interact with social machines.

As a result, it is important to understand how the FAE may impact perceptions of a social robot when the robot engages in popular or unpopular behavior. These findings will have implications for how humans understand the causes of social robots' behavior and assign blame or credit for their activities, which is increasingly relevant in contexts including emergency/crisis, healthcare, education, retail, and legal. In short, how will people assign the cause of a robot's behavior in relation to how they do so for other humans? More specifically, to closely replicate the experimental research on FAE in human interaction (Forgas, 1998), we will focus on a situation in which a robot or human expresses the popular or unpopular position on a topic of social importance. This design falls within the attributed attitude paradigm of research investigating the correspondence bias (Jones and Harris, 1967). Although application of the CASA paradigm would suggest people will demonstrate similarity in their attributional processes of human and robot behavior, the observed differences in people's responses may indicate differences. The traditional procedure for carrying out research within the CASA framework entails 1) selection of a theory or phenomenon observed in human interaction, 2) adding humanlike cues to a robot, 3) substituting the robot for a human actor, and 4) determining whether the same social rule applies (Nass et al., 1994). To also allow for identification of more granular potential differences in how people respond to robots, the present study modifies and extends the procedure to include a human-to-human comparison group. We offer the following research questions:

RQ1: Will participants attribute the cause of an agent's (social robot or human) behavior to disposition or situational factors?

RQ2: How will the nature of an agent's behavior (popular or unpopular) influence attributions and impressions?

3 MATERIALS AND METHODS

3.1 Participants

The sample included 267 U.S. American adults recruited and compensated US $2.00 through Amazon's Mechanical Turk. Participants who 1) failed the audio test, 2) failed the speech-topic attention check or 3) reported non-normative attitudes toward the topic (opposed legalization of medicinal marijuana) were excluded from analysis, leaving 231 participants. Their average age was 43.32 years (*SD* = 11.36, *MD* = 40, *range* = 24–71). Slightly over half identified as male (51%, *n* = 118), followed by female (48%, *n* = 110), those who selected "prefer to not answer" (0.9%, *n* = 2), and gender fluid (0.4%, *n* = 1). Predominantly, participants identified as White (79%, *n* = 183), followed by Black or African-American (7%, *n* = 16), Hispanic or Latino/a/x (5%, *n* = 12), Asian or Pacific Islander (5%, *n* = 12), bi- or multi-racial (3%, *n* = 7), and one person (0.4%) selected "prefer to not answer". Most had a Bachelor's degree or higher (60%, *n* = 138).

3.2 Procedures

Procedures entailed a modified replication of Forgas' (1998) experiments investigating the correspondence bias by examining the degree to which people attributed a person's message behavior to their "true attitudes" about the topic when that behavior was popular (normative, and therefore expected) or unpopular, and chosen or coerced (Forgas, 1998). Additionally, Forgas manipulated the mood of participants as happy or sad to determine the influence of mood on attributional judgements. Participants were asked to read an essay forwarding either a popular or unpopular position on the topic of French nuclear testing in the South Pacific, which was framed as either the chosen stance of the author or an assigned/coerced stance. Then, participants were asked to consider whether the essay represented the true attitude of its writer, to indicate their degree of confidence in that attribution, and to give their impressions of the essay writer. In the present study, we replicated the basic design with four modifications: 1) manipulation of the agent as human or robot, 2) use of a more contemporary topic (medicalization of marijuana), 3) speeches versus essays, and 4) measured and statistically controlled for mood rather than manipulated mood.

Upon securing Institutional Review Board approval and obtaining informed consent, we conducted a 2 (agent: human vs. robot) X 2 (behavior: popular vs unpopular) X 2 (choice: chosen vs. coerced) between-subjects online video experiment, which was introduced to participants as a "social perception study." After completing an audio check, participants were asked to rate their current affective/mood state. Next, participants were randomly assigned to view one of eight experimental conditions involving a 1-min video containing a persuasive appeal by a human or a robot, in which the agent advocated for or against legalizing medical marijuana (operationalizing popular vs.

unpopular behavior), with the position stipulated as either freely chosen by or assigned to the speaker. As a manipulation and attention check, participants were asked to report the speaker's stated position in the video before progressing to the rating tasks. Next, they were asked a series of questions assessed along 7-point semantic differential scales to ascertain 1) inferences of the speakers' "true attitudes" toward legalizing medical marijuana, 2) confidence in their attributed attitude ratings, and 3) interpersonal impressions of the speaker. Finally, they were asked to report their own attitudes toward the legalization of medical marijuana, to offer any open-ended comments, and to provide demographic information.

3.3 Mood Check

Prior to the experimental task, participant mood was self-assessed with two (1:7) semantic differential items rating current mood as sad:happy and bad:good. Answers were highly related [r (228) = 0.92, $p < 0.001$] and therefore summed to create a single mood score, *alpha* = 0.96, *item M* = 5.98, *SD* = 1.40.

3.4 Attribution Task

Participants were asked to "carefully watch a 1-min persuasive speech written by this (person/robot)," who, they were informed, either chose to take this stance on the issue (choice) or was assigned to do so (coerced). Next, they were asked to answer a series of questions about the speaker. The speeches dealt with the familiar and salient topic of the legalization of marijuana for medical purposes. There is a strongly preferred normative position on the issue, with 91% of U.S. Americans in favor of legalization for medical use [59% for medical and recreational +32% for medical use only; (Daniller, 2019)]. The speeches persuading for and against legal marijuana (popular and unpopular behavior, respectively) were identical except for single phrases or words substituted to reverse the sentiment and meaning of the two parallel conditions. For example, "Medical marijuana should (should not) be legal," "Legalizing medical marijuana is (is not) in the public's best interest" and "Legal medical marijuana will (will not) be effectively regulated for consumers." The overall position forwarded in each speech was clearly and strongly for or against legalization.

3.5 Agent Manipulation

For the human agent conditions, a graduate research assistant unknown to participants delivered the 1-min persuasive appeal for and against legalization. The robot agent was Softbank's Pepper humanoid robot, which was programmed to deliver the same scripted speeches with a matching rate of speech and comparable animacy of gestures and movement. For both the human and robot conditions, the video frame included the face and upper body in front of a light-colored backdrop.

3.6 Dependent Variables

After watching the speeches, participants rated the speaker along a series of 7-point bipolar scales, which assessed 1) perceptions of the speaker's "real attitudes" toward the issues ("What do you think the speaker truly believes about legalizing medicinal marijuana?" Supports it–Opposes it), 2) levels of confidence in attributed attitudes ("How confident are you in knowing what the speaker truly believes about the issues?" Confident–Not Confident), and 3) global impressions of the speaker (Dislikable–Likable; Unpopular–Popular: Unintelligent–Intelligent; Incompetent–Competent; Untrustworthy–Trustworthy; Inexpert–Expert; Uncaring–Caring; Unsimilar–Similar), with items similar to those used to in previous studies of the correspondence bias e.g., (Forgas, 1998).

TABLE 1 | Means and standard deviations for attributed attitudes (1–7; opposes it:supports it).

		Behavior					
		Popular		Unpopular		Total	
Agent	Choice	M	SD	M	SD	M	SD
Human	Chosen	5.72	1.12	3.81	1.94	4.86	1.83
	Coerced	5.31	1.04	4.00	1.94	4.81	1.57
Robot	Chosen	5.91	1.23	2.62	1.60	4.29	2.18
	Coerced	5.63	1.35	2.48	1.35	4.08	2.08
Total	Chosen	5.82	1.21	3.16	1.84	4.56	2.03
	Coerced	5.47	1.21	3.06	1.75	4.41	1.90

3.7 Attitude Assessment

Participants' attitudes toward the issue of legalizing medical marijuana was also assessed. Approximately 92% of the sample supported the position that "medical marijuana should be legal," which indicated the strong popularity of the pro-legalization speech stances. As noted above, potential participants who opposed the legalization of medical marijuana were subsequently excluded from analysis to ensure that pro-legalization speeches operationalized "popular" behavior [i.e., a strongly preferred normative and therefore probabilistic opinion; (Jones and Harris, 1967)].

4 RESULTS

4.1 Mood

Participants' mood states at the beginning of the experiment were statistically controlled as a covariate in all analyses because mood has been found to influence the degree to which judges demonstrate the correspondence bias. Specifically, happy mood enhanced and sad mood lessened dispositional attributions of coerced unpopular behavior (Forgas, 1998).

4.2 Attribution of Attitude to the Speaker

A three-way analysis of covariance (ANCOVA) evaluated the effects of agent (human vs. robot), behavior (popular vs. unpopular), and choice (chosen vs coerced) on attributed attitudes while controlling for mood. See **Table 1** for means and standard deviations.

As expected, in a significant main effect of behavior, agents that expressed popular versus unpopular positions were judged to hold significantly different attitudes about the issue (M = 5.65 vs. 3.11), F (1, 220) = 152.98, $p < 0.001$, *partial eta squared* = 0.41. There was also a significant main effect of agent with stronger pro-legalization attitudes attributed to the human versus robot

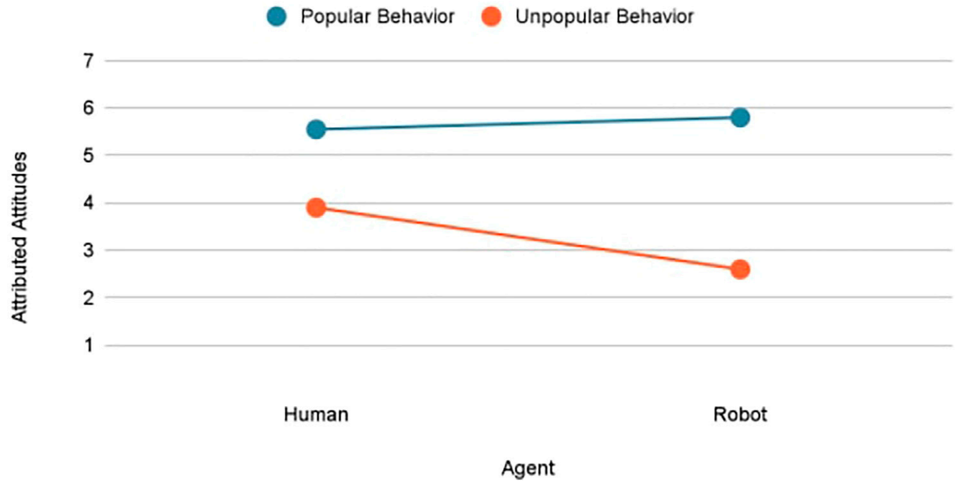

FIGURE 1 | Interaction effect of agent and behavior on attributed attitudes.

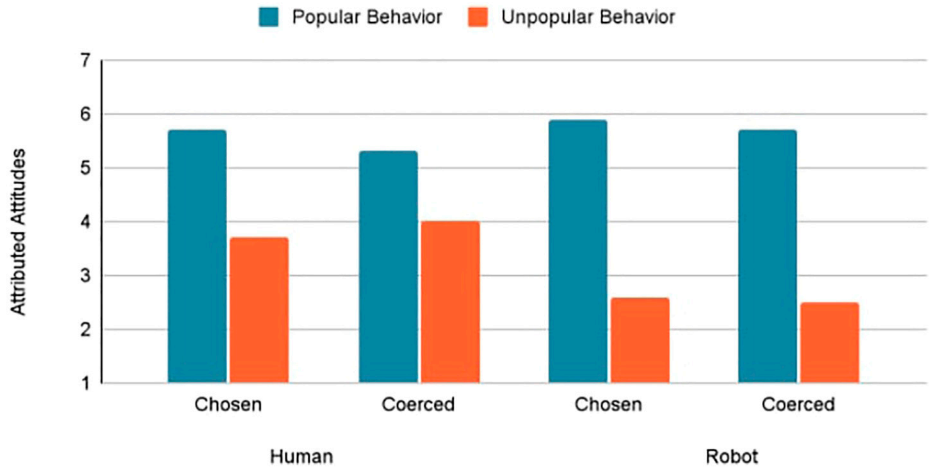

FIGURE 2 | Simple main effects of choice and behavior on attitudes attributed to agents.

(M = 4.84 vs. 4.19), F (1, 220) = 7.39, $p < 0.01$, *partial eta squared* = 0.03. The two-way interaction between behavior and choice was not significant, F (1, 220) = 0.84, p = 0.36, *partial eta squared* = 0.00. Differing, topic-congruent attitudes were attributed to agents that expressed popular versus unpopular positions regardless of whether their stances were chosen or coerced, establishing that judges demonstrated a correspondence bias, or FAE, in attributing attitudes.

As shown in **Figure 1**, there was a significant interaction between behavior (popular vs. unpopular) and agent (human vs. robot), F (1, 220) = 16.15, $p < 0.001$, *partial eta squared* = 0.07. Analysis of simple main effects showed that different attitudes were attributed to agents expressing popular versus unpopular positions in both the human (M = 5.52 vs. 3.89) and robot (M = 5.78 vs. 2.56) conditions.

As depicted in **Figure 2**, agent type had no marked influence on attributions of popular behavior (M = 5.52 vs. 5.78). With unpopular behavior, however, the judges inferred the robot to have a stronger topic-congruent attitude compared to the human (M = 2.56 vs. 3.89).

Results confirmed that judges made correspondent inferences of both human [F (1, 100) = 27.12, $p < 0.001$, *partial eta squared* = 0.21] and robot agents [F (1, 119) = 161.94, $p < 0.001$, *partial eta squared* = 0.58], by assuming that their true attitudes aligned with their expressed positions. However, the effect size of behavior (popular vs. unpopular) on attributed attitudes was substantially larger for robots than humans. Judges drew a stronger unit relation between the agent and its behavior when evaluating the robot, as further demonstrated by linear regressions treating attributed attitudes as criterion and behavior valence as predictor in human and robot conditions. When judging humans, behavior valence was a significant predictor of attributed attitudes, *Beta* = –0.48, t (104) = –5.66, $p < 0.001$, and explained significant variance in attributed

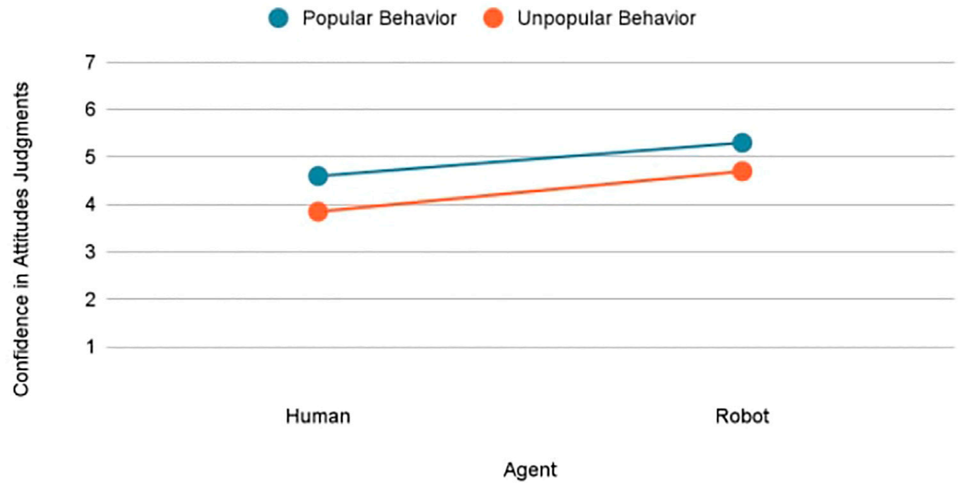

FIGURE 3 | Effects of agents and behavior on confidence in attitude attribution.

attitudes, *adjusted* r^2 = 0.23, F (1, 104) = 32.02, p < 0.001. However, behavior valence was a stronger predictor of attitudes attributed to robots [*Beta* = –0.76, t (122) = -12.95, p < 0.001] and produced a larger effect size [*adjusted* r^2 = 0.58, F (1, 122) = 167.62, p < 0.001]. The relatively stronger correspondence bias toward robots was driven by the greater dispositional culpability attributed to robots engaging in unpopular behavior (anti-legalization stance), whether freely chosen or coerced.

4.3 Confidence in Attitude Judgments

Confidence ratings for attitude judgments were analyzed to assess any awareness by judges of their attributional limitations (**Figure 3**). Choice had no significant influence on confidence in attributions [F (1, 220) = 1.71, p = 0.193, *partial eta squared* = 0.01]; participants felt equally confident in their attitude attributions of speakers whose positions were chosen versus coerced. Both agent [F (1, 220) = 6.02, p = 0.015, *partial eta squared* = 0.03] and behavior [F (1, 220) = 5.04, p = 0.026, *partial eta squared* = 0.02] had a main effect on confidence. Judges reported greater confidence in their attributions of popular versus unpopular positions (M = 4.94 vs. 4.42) and of robot versus human agents (M = 4.95 vs. 4.42). Judges drew a stronger unit relation between the robot agent and its actions, and also felt greater confidence in their judgments of the robot's attitudes.

4.4 Impressions

Impression judgments on the eight bipolar scales were factor analyzed. Visual inspection of the scree plot and consideration of Eigenvalues > 1.00 supported treatment as unidimensional (Eigenvalue = 5.27; 65.86% variance; highest loading item = unlikable:likable). Therefore, we summed the items to form the impressions dependent variable, *alpha* = 0.92 (*item M* = 5.21, *SD* = 1.45). The effects of agent, behavior, and choice on impressions were assessed with a three-way ANCOVA, again controlling for mood. **Table 2** for means and standard deviations.

TABLE 2 | Means and standard deviations for impressions (1–7; Negative: Positive).

		Behavior					
		Popular		Unpopular		Total	
Agent	Choice	M	SD	M	SD	M	SD
Human	Chosen	6.10	1.01	4.93	1.58	5.56	1.41
	Coerced	5.74	0.94	5.19	1.29	5.53	1.11
Robot	Chosen	5.78	1.26	3.91	1.33	4.86	1.59
	Coerced	5.29	1.58	4.71	1.29	5.00	1.46
Total	Chosen	5.93	1.14	4.38	1.52	5.19	1.54
	Coerced	5.51	1.32	4.89	1.30	5.29	1.34

There was a significant main effect of agent [F (1, 221) = 8.75, p = 0.003, *partial eta squared* = 0.04] and of behavior [F (1, 221) = 38.26, p < 0.001, *partial eta squared* = 0.15] with more favorable ratings of humans versus robots (M = 5.54 vs. 4.92) and of agents expressing popular versus unpopular positions (M = 5.73 vs. 4.60). There was no main effect of choice [F (1, 221) = 0.28, p = 0.594, *partial eta squared* = 0.001]. However, choice condition and behavior interacted to influence impressions of the agent [F (1, 220) = 7.78, p = 0.006, *partial eta squared* = 0.03]. Because judges formed different impressions of human and robot agents, simple main effects were examined separately for agent conditions (**Figure 4**).

4.4.1 Impressions of Human Agent

The human was rated more favorably when taking the popular versus unpopular stance, F (1, 101) = 18.40, p < 0.001, *partial eta squared* = 0.15, M = 5.93 vs. 5.03). There was no significant effect of choice [F (1, 101) = 0.052, p = 0.820, *partial eta squared* = 0.001] or interaction effect of choice and behavior [F (1, 101) = 1.45, p = 0.231, *partial eta squared* = 0.014] on interpersonal impressions.

4.4.2 Impressions of Robot Agent

Choice and behavior interacted to influence interpersonal impressions of the robot, F (1, 119) = 6.72, p = 0.011, *partial eta squared* = 0.05. Robots expressing the popular position

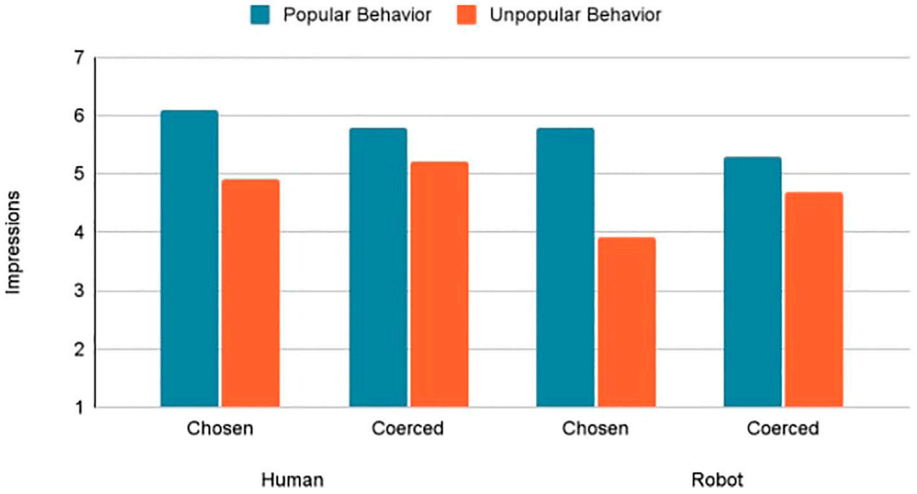

FIGURE 4 | Simple main effects of choice and behavior on impressions of agents.

garnered the same impressions whether the position was chosen or coerced, *F* (1, 60) = 1.67, *p* = 0.201, *partial eta squared* = 0.019; *M* = 5.78 vs. 5.29. In contrast, a robot expressing the unpopular position was perceived more negatively than a robot coerced to do so (*M* = 3.90 vs. 4.71), *F* (1, 58) = 5.60, *p* = 0.021, *partial eta squared* = 0.09. When the robot's behavior was freely chosen, the popular stance led to more favorable interpersonal impressions than the unpopular stance, *F* (1, 62) = 32.34, *p* < 0.001, *partial eta squared* = 0.34, *M* = (5.78 vs. 3.91). When the robot's behavior was coerced, there was no significant difference in the interpersonal impressions formed of popular versus unpopular behavior, *F* (1, 56) = 2.28, *p* = 0.136, *partial eta squared* = 0.04, *M* = 5.29 vs. 4.71). Judges were more generous in their impressions of the robot when its unpopular behavior was coerced rather than chosen; a tendency not displayed when forming impressions of the human agent. Although judges formed different impressions of the robot that chose its position, the direction of coerced position had no marked influence on impressions.

5 DISCUSSION

The correspondence bias (FAE) has been thoroughly tested with people, but not with HRI. In general, people tend to overemphasize dispositional explanations for behaviors seen in others and, at the same time, under-emphasize features of the situation (Pak et al., 2020). Because a social robot's behavior is completely determined by its design, programming, and humans behind the scenes, it is essential to know if people will still commit the correspondence bias for robot behavior. These findings have implications for assigning credit or blame to a social robot's behaviors. In this section, we will summarize the results, discuss implications, and offer limitations and directions for future research.

5.1 Summary of Results

Research question 1 asked if participants would attribute the cause of an agent's (social robot or human) behavior to disposition or to situational factors. Participants exhibited the correspondence bias (FAE) toward both human and robot agents by assuming their behavior corresponded to their underlying attitudes (a dispositional attribution) even when their behavior was clearly assigned (a situational cause). However, their dispositional correspondent inferences were stronger for the robot than for the human. In other words, judges of the robot drew a stronger unit relation between the actor and its actions, as evidenced by the larger effect size of popular or unpopular behavior on attributed attitudes for the robot. With unpopular behavior, specifically, judges held the robot more dispositionally culpable than the human. Judges also felt greater confidence in their judgments of the robot's true attitudes compared to the human.

Research question 2 asked if the nature of the agent's behavior as popular or unpopular would influence causal attributions and global impressions. The relatively stronger correspondence bias toward robots was driven by the greater dispositional culpability attributed to robots committing unpopular behavior, whether freely chosen or coerced. Participants generally formed more favorable impressions of human versus robot agents and popular behavior versus unpopular behavior. Humans were rated more favorably for popular behavior than for unpopular behavior, regardless of whether they chose or were assigned the behavior.

When forming impressions of robots, there were some differences. For robots committing popular behavior, the same attitudes were attributed to them whether they chose or were assigned. However, judges were more generous in their impressions of the robot when its unpopular behavior was coerced rather than chosen, a tendency not displayed when forming impressions of the human agent. Although judges formed different impressions of the robot that chose to commit popular or unpopular behavior, coerced behavior type had no marked influence on impressions. Paradoxically, people held the robot more dispositionally responsible for its forced unpopular behavior than its chosen unpopular behavior, but were also more generous in their global impressions of the robot when its unpopular behavior was forced. Although judges formed different and valence-congruent impressions of the robots that chose popular or unpopular behavior, the impressions they formed of robots coerced to commit popular or unpopular behavior did not differ.

5.2 Implications

First, there were similarities in how participants made causal attributions about robot and human behavior. They made correspondent inferences for both, attributing the cause of behavior to the agent's disposition even when the agent was coerced to do it. This may support the CASA Paradigm (Reeves and Nass, 1996) by showing similarities in how we treat social robots and people, consistent with prior research drawing parallels in terms of robot mind ascription, intention, goals, and so on. However, the differences between how participants judged humans and robots are perhaps more interesting and important. At the broadest level, these differences in how a classic social psychology finding applied to robots versus humans adds to a small set of studies challenging the notion that people necessarily interpret and react to social robots as they do to other humans. For instance, in an HRI replication of The Milgram Shock Experiment (Bartneck et al., 2005), found that every participant was willing to administer to a robot the highest voltage shock, whereas 60% of participants in the original study refused to use the maximum setting on another human. Furthermore, there are documented differences in the expectations for interaction people hold of social robots versus humans (Spence et al., 2014; Edwards et al., 2016; Edwards et al., 2019) and in their ontological understandings of these agents (Kahn et al., 2011; Edwards, 2018). Results of this experiment are also consistent with the idea that people view robots as unique from humans on dimensions including social presence, agency, free will, status, and capacity for suffering, which may lead them to develop and apply media-centric scripts developed specifically for cognition and behavior toward social robots (Gambino et al., 2020). Although both computer-based technologies and humans may be social actors (CASA), they are not necessarily seen as the same type of social actor.

The question becomes, what is the significance of the specific differences observed in this experiment: 1) that there were stronger dispositional correspondent inferences (stronger actor/agent conflation) for robots than for humans, 2) that people were more certain about a robot's "true disposition" than a human's, and 3) that people uncoupled attributed attitudes from global impressions to a greater degree for robots? Satisfying answers will depend upon why people (appeared to) not only commit the fundamental attribution error with robots–which are machines logically understood to operate without interior "dispositions" like personality, attitudes, beliefs, and feeling–but also to commit it to a greater degree and with greater certainty then they did with humans. At first glance, these causal inferences of robot behavior may appear to be a mistake or error akin to the one people make in judging one another.

However, there are three problems with calling the observed results an instance of Fundamental Attribution Error (FAE). The first two arise from cross-application of criticism surrounding human FAE studies using attributed attitude paradigms: 1) the judge never really knows whether the coerced actor actually agrees or disagrees with the direction of their forced action, which means a dispositional attribution is not necessarily incorrect/erroneous and 2) correspondent inferences in which an actor is presumed to possess action-congruent attitudes do not necessarily mean that the central underlying premise of FAE—that people routinely overemphasize dispositional and underemphasize situational causes of behavior—has been supported. These critiques have resulted in a preference for the terms "correspondence bias" or "dispositional correspondent inferences" over FAE when there is no direct test of Situation Theory (S-Theory) awareness and its role in attribution processes (Gawronski, 2004). In the case of robots, there is a third and obvious reason to hesitate to apply the term "error" to a tendency to infer that a robot's behavior corresponds to its disposition: Logically, it does not seem possible that robots, as programmed machines, hold true dispositions, beliefs, or attitudes that are incongruent with their actions. This is because beliefs and attitudes are widely understood to require inward experiential aspects or subjectivity of thought that does not characterize present robots.

Therefore, viewing the results through the lens of the correspondence bias is more fruitful because it removes both the evaluative aspect of whether people are "right" or "wrong" to conflate a robot agent and its actions and the necessity of linking observed effects to a broad and pervasive underestimation of situational influence, to center only on whether people bend toward disposition-situational convergence. Now the issue remains of how to interpret the relatively stronger and more confident correspondence bias people exhibited toward social robots. As discussed by Gawronski (2004), the correspondence bias may arise from a number of different processes involving how people apply causal theories about the role of situation on behavior (S-Theory). These include 1) lack of S-Theory (when there is no awareness of or there is disagreement with the premise that situational factors constrain behavior), 2) failed application of S-Theory (when there is knowledge of and belief in S-Theory adequacy, but people are unmotivated, lack cognitive capacity, or have inferential goals which result in failure to correct dispositional attribution bias), 3) deliberate neglect of S-Theory (when S-Theory is deemed irrelevant because observed behavior seems highly diagnostic irrespective of situational forces, as in cases of morality and performance ability), and 4) biasing application of S-Theory (when S-Theory is applied in a manner that amplifies rather than attenuates correspondent dispositional inferences) (Gawronski, 2004).

This fourth and final cause of correspondence bias—biasing application of S-Theory—seems especially relevant to understanding why people may make stronger correspondent dispositional inferences for robots than for other humans. The "over-" or biasing application of S-Theory (where "over" implies an attributional effect and not a normative or judgmental inadequacy) may occur in cases in which people understand that behavior is constrained by situational factors, are aware of present situational factors (e.g., whether the behavior was freely chosen or assigned, and the nature of the agent), have the capacity and motivation to apply S-theory, then do so to such a high degree that it appears as if they have totally ignored the causal role of situational factors (Gawronski et al., 2002). For example, people may disambiguate ambiguous human behavior by defining disposition completely in terms of the situation; Ambiguous behavior has been attributed to dispositional anxiety because the situation was perceived as anxiety-inducing (Snyder and Frankel, 1976).

Theoretically, people's ideas about what robots are, how they work, and how they compare to humans could also lead to a biasing application of S-Theory. To the degree that robots are understood as programmed and controlled by humans, the situation may become salient to the degree it is considered completely determinative of and the same thing as disposition (they are programmed, hence their behavior literally is their personality/attitude/disposition). Ironically, this strong or complete application of S-Theory would appear in the data as heightened dispositional inference because participants would presume alignment between the robot's behavior and its true attitudes or personality. In reality, this pattern of findings may simply reflect participants' tendency to conflate an agent whose nature is to lack independent, interior life with its situationally determined actions.

Perhaps most significantly for theorizing HRI, this possible explanation prompts serious consideration of the idea that people may use different causal attribution processes to display a correspondence bias with robots than they do with other humans, even under the same circumstances. Both the stronger and more certain unit relation participants drew between a robot actor and its actions and the looser relationship they displayed between attributed attitudes and general impressions of the robot (i.e., the greater impression-related generosity for robots coerced to do unpopular things compared to humans) compel further investigation into whether unique perceptual patterns and theoretical mechanisms underlie causal inferences of robot behavior. Naturally, people's causal theories about the role of situation on behavior (S-Theory) may be different for robots and human beings because people perceive them to be ontologically distinct (Kahn et al., 2011; Edwards, 2018).

The FAE, from which correspondence bias research derived, has been called the conceptual bedrock of the field of social psychology, which rests on the assumption that we tend to see others as internally motivated and responsible for their own behavior (Ross, 1977). Drawing a distinction between personality and situation is meaningful when making sense of other humans, and it appears to factor prominently in the dispositional correspondent inferences we tend to make of one another. But with robots, the similar-appearing, but stronger correspondence bias demonstrated by participants could arise from a different psychology altogether, and one more akin to the analytical/logical behaviorism which equates behavioral and mental tendency. Viewed from this lens, much of our descriptive vocabulary for human beings—mind, personality, intention, disposition, attitude, belief—may still be productively transferred to robots, but meant in a different sense [see, e.g., (De Graaf and Malle, 2019)]. Thellman et al. (2017) suggest a similar explanation of their finding that when asked explicitly, people rated goals and dispositions as a more plausible cause of behavior when the actor was human: "This raises the question whether people think of robots as less likely to have dispositions in the human sense, or as having less stable dispositions as humans, or whether people see robot dispositions as less efficacious in causing behavior than human dispositions" (Thellman et al., 2017, p. 11).

Our participants readily attributed to the robot a "true" or "real" attitude and they inferred the nature of that attitude heavily from observed behavior. However, is a robot attitude the same thing as a human attitude (see Nilsson, 2014, on robot "belief")? Or, is the latter understood to be held (and therefore possibly concealed or subordinated), while the former is purely beheld (manifest, observed, perceived through sight or apprehension), rendering the causal distinction between an agent and its action unhelpful or illogical in the case of robots?

In other words, might people be social psychologists when it comes to other humans and behavioral psychologists when it comes to robots? For commentary on the application of behaviorist principles to robots, see: Sætra (2021); Danaher (2019).

Naturally, working out the fruitfulness of the paths of inquiry suggested above will require asking people what they think about the meaning of attitudes, beliefs, or personality (and situation) in the context of robots, and observing their language and behavior both *in situ* and in experiments designed specifically to test alternative explanations for a correspondence bias (or "agent-action conflation bias") in HRI and to chart the boundaries of when, where, why, and how it may converge or diverge from human-centric causal inference processes.

In terms of methodological implications for the study of HRI, this research demonstrates the value of including within HRI experiments a human-human condition. Classically, research undertaken within the CASA framework encourages choosing a social science finding (theory and method) that applies to human interaction, replicating the research while substituting a robot/computer for a human actor in the statement of theory and design, providing the robot with human-linked characteristics, and determining whether and to what degree the same social rule still applies (Nass et al., 1994). We argue that including a human-human comparison group offers three advantages to the traditional methodology: 1) it tests again the applicability of the theory to human behavior, which is important given recent replication and reproducibility difficulties in social fields (Maxwell et al., 2015), 2) allows for the identification of both similarities and differences in HHI and HRI (including effect magnitudes) without relying on comparisons between dissimilar datasets and samples, and 3) opens examination of the possibility that even patterns of similarity in HHI and HRI may manifest for a different reason than the mindless application of human social scripts to interactions with robots (Edwards et al., 2019; Gambino et al., 2020; Fortunati and Edwards, 2021). This latter point is especially crucial because the original procedure to conduct CASA research is not sensitive to the potential operation of different theoretical mechanisms responsible for similar observational endpoints. Had we not included a human condition in this experiment, the results would have appeared only to generally mirror a tendency found in human interaction (Forgas, 1998); to suggest people also overemphasize personality at the expense of situational consideration when explaining robot behavior) and left unaddressed questions including "Are we certain the correspondence bias would be replicated with humans today, in this historical and cultural context?" "Are there any differences in how our participants would have evaluated human beings performing the same actions in the same situation?" and "Do any differences, large or small, suggest the possibility that even observed similarities warrant interpretive scrutiny?".

5.3 Future Research

The current study demonstrates that the correspondence bias extends to human-robot interactions. We do not know what factors influence the situational and dispositional attributions people make about robots. Do people over-apply situational theory to robots? In other words, how can bias attenuation occur in an interaction? Identifying future research needs to examine, through experimental design, why exactly people appear to make stronger correspondent inferences for robots than humans and how that will translate to the assignment of credit, blame, moral agency, and moral patiency. Additionally, future research needs to examine what factors may enhance or attenuate correspondent inferences.

People have an anthropocentric bias about conversations in that they expect to speak with a human and not a machine partner. In these studies, people report lower liking for social robots and have greater uncertainty about the interaction (Spence et al., 2014; Edwards et al., 2016; Edwards et al., 2019). Do these findings impact potential attributional errors with social robots? And if so, what can be done to attenuate them? Does the greater uncertainty cause the over-application? Aspects of the robot, including morphology, scripting, interaction modality, and interaction history, should be explored for potential effects on causal attributions of its behavior. Future research needs to explore why people held the robot more dispositionally responsible than the human and why they felt greater confidence in their judgments of the robot's attitudes than the human actor.

Third, how exactly is responsibility handled differently with robots than humans? Because participants were relatively more kind in their reported impressions of the robot when its bad behavior was coerced (not so for the human agent), we need future research to examine how responsibility for decision-making will occur. Previous research has demonstrated that even when participants are given transparent details about robot behaviors and drives, they thought the robot was thinking more (Wortham et al., 2017). Although it is possible that the meaning of robot "thinking" shifted following explanations of how the robot functioned. We suspect that interpersonal relationship dimensions will come into play. If we have a relationship with a social robot, do we offer more responsibility for decision-making to the robot? We certainly do with people, and it stands to reason that relationships will make a difference in HRI. In the current study, the exposure time was the same for each condition and yet the robot was held more dispositionally responsible. Future research needs to examine if relationship differences can attenuate these differences.

Finally, it is possible that the video stimulus was not as "real-world" as a study with face-to-face embodied presence with the robot. Furthermore, the scenario was hypothetical and pertained to a single, short speech. Potentially, attribution processes play out differently following longer-term, real-world observation of robot behavior, and could differ when evaluating message behavior versus other types. Future research should replicate this study in a live interaction. Being in the room with a social robot might cause a differing correspondence bias than simply watching one on a video. Issues such as social presence (Short et al., 1976) might impact these judgments.

6 CONCLUSION

This study demonstrates that people do exhibit the correspondence bias with social robots. This experiment shows a stronger correspondence bias toward social robots than humans, or the tendency to conflate an agent and its actions into a single categorical unit. Especially in the case of unpopular behavior, judges inferred the robot had more congruent underlying attitudes than the human. The tendency to believe that what people do reflects who they are may be magnified in HRI to the degree that people think what robots do is who they are. People held robots more dispositionally responsible for their unpopular behavior, and people were more confident in their attributions of a robot than human attitudes. Although participants attributed behavior congruent beliefs to robots as they did to other humans, they perhaps did not attribute the possibility of true thoughts incongruent with their actions. As such, we may be social psychologists when interpreting other people and behaviorists when interpreting robots.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The participants provided online informed consent to participate in the study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This project was funded through research funds provided by the Communication and Social Robotics Labs at the School of Communication at Western Michigan University.

REFERENCES

Banks, J., Edwards, A. P., and Westerman, D. (2021). The Space between: Nature and Machine Heuristics in Evaluations of Organisms, Cyborgs, and Robots. *Cyberpsychology, Behav. Soc. Networking* 24, 324–331. doi:10.1089/cyber.2020.0165

Banks, J. (2020). Theory of Mind in Social Robots: Replication of Five Established Human Tests. *Int. J. Soc. Robotics* 12, 403–414. doi:10.1007/s12369-019-00588-x

Barrett, J. L. (2007). Cognitive Science of Religion: What Is it and Why Is it? *Religion Compass* 1, 768–786. doi:10.1111/j.1749-8171.2007.00042.x

Barrett, J. L. (2000). Exploring the Natural Foundations of Religion. *Trends Cognitive Sciences* 4, 29–34. doi:10.1016/s1364-6613(99)01419-9

Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., and Kennsuke, K. (2005). "A Cross-Cultural Study on Attitudes towards Robots," in HCI International Conference, Las Vegas, NV, July 2005.

Berry, D. M., and Frederickson, J. (2015). The Postdigital Constellation. *J. Integrated Soc. Sci.* 5, 44–57. doi:10.1057/9781137437204_4

Branyon, J., and Pak, R. (2015). "Investigating Older Adults' Trust, Causal Attributions, and Perception of Capabilities in Robots as a Function of Robot Appearance, Task, and Reliability," in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA (Los Angeles, CA): SAGE Publications Sage CA) v.59(1), 1550–1554. doi:10.1177/1541931215591335

Ciardo, F., Beyer, F., De Tommaso, D., and Wykowska, A. (2020). Attribution of Intentional agency towards Robots Reduces One's Own Sense of agency. *Cognition* 194, 104109. doi:10.1016/j.cognition.2019.104109

Danaher, J. (2019). The Philosophical Case for Robot friendship. *J. Posthuman Stud.* 3, 5–24. doi:10.5325/jpoststud.3.1.0005

Daniller, A. (2019). *Two-thirds of Americans Support Marijuana Legalization*. Pew Research Center.

De Graaf, M. M., and Malle, B. F. (2017). "How People Explain Action (And Autonomous Intelligent Systems Should Too)," in 2017 AAAI Fall Symposium Series, October 2017.

De Graaf, M. M., and Malle, B. F. (2019). "People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences," in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (IEEE), 239–248. doi:10.1109/hri.2019.8673308

Edwards, A. (2018). "Animals, Humans, and Machines: Interactive Implications of Ontological Classification," in *Human-machine Communication: Rethinking Communication Technology and ourselves* (New York, NY: Peter Lang), 29–50.

Edwards, A., Edwards, C., and Gambino, A. (2020). The Social Pragmatics of Communication with Social Robots: Effects of Robot Message Design Logic in a Regulative Context. *Int. J. Soc. Robotics* 12, 945–957. doi:10.1007/s12369-019-00538-7

Edwards, A., Edwards, C., Westerman, D., and Spence, P. R. (2019). Initial Expectations, Interactions, and beyond with Social Robots. *Comput. Hum. Behav.* 90, 308–314. doi:10.1016/j.chb.2018.08.042

Edwards, C., Edwards, A., Spence, P. R., and Westerman, D. (2016). Initial Interaction Expectations with Robots: Testing the Human-To-Human Interaction Script. *Commun. Stud.* 67, 227–238. doi:10.1080/10510974.2015.1121899

Forgas, J. P. (1998). On Being Happy and Mistaken: Mood Effects on the Fundamental Attribution Error. *J. Personal. Soc. Psychol.* 75, 318–331. doi:10.1037/0022-3514.75.2.318

Fortunati, L., and Edwards, A. P. (2021). Moving Ahead with Human-Machine Communication. *Human-Machine Commun.* 2, 1. doi:10.30658/hmc.2.1

Gambino, A., Fox, J., and Ratan, R. A. (2020). Building a Stronger Casa: Extending the Computers Are Social Actors Paradigm. *Human-Machine Commun.* 1, 5. doi:10.30658/hmc.1.5

Gawronski, B., Alshut, E., Grafe, J., Nespethal, J., Ruhmland, A., and Schulz, L. (2002). Prozesse der Urteilsbildung über bekannte und unbekannte Personen. *Z. für Sozialpsychologie* 33, 25–34. doi:10.1024//0044-3514.33.1.25

Gawronski, B. (2004). Theory-based Bias Correction in Dispositional Inference: The Fundamental Attribution Error Is Dead, Long Live the Correspondence Bias. *Eur. Rev. Soc. Psychol.* 15, 183–217. doi:10.1080/10463280440000026

Gilbert, D. T., and Jones, E. E. (1986). Perceiver-induced Constraint: Interpretations of Self-Generated Reality. *J. Personal. Soc. Psychol.* 50, 269–280. doi:10.1037/0022-3514.50.2.269

Haring, K. S., Watanabe, K., Velonaki, M., Tossell, C. C., and Finomore, V. (2018). FFAB-the Form Function Attribution Bias in Human-Robot Interaction. *IEEE Trans. Cogn. Dev. Syst.* 10, 843–851. doi:10.1109/tcds.2018.2851569

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., and Sagerer, G. (2009). "Understanding Social Robots," in 2009 Second International Conferences on Advances in Computer-Human Interactions, Cancun, Mexico, February 2009 (IEEE), 169–174. doi:10.1109/achi.2009.51

Heider, F., and Simmel, M. (1944). An Experimental Study of Apparent Behavior. *Am. J. Psychol.* 57, 243–259. doi:10.2307/1416950

Heider, F. (1958). *The Naive Analysis of Action*. New York, NY: Wiley.

Jones, E. E., and Harris, V. A. (1967). The Attribution of Attitudes. *J. Exp. Soc. Psychol.* 3, 1–24. doi:10.1016/0022-1031(67)90034-0

Jones, E. E. (1979). The Rocky Road from Acts to Dispositions. *Am. Psychol.* 34, 107–117. doi:10.1037/0003-066x.34.2.107

Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., et al. (2011). "The New Ontological Category Hypothesis in Human-Robot Interaction," in 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Lausanne, Switzerland, March 2011 (IEEE), 159–160. doi:10.1145/1957656.1957710

Langdridge, D., and Butt, T. (2004). The Fundamental Attribution Error: A Phenomenological Critique. *Br. J. Soc. Psychol.* 43, 357–369. doi:10.1348/0144666042037962

Martini, M. C., Gonzalez, C. A., and Wiese, E. (2016). Seeing Minds in Others - Can Agents with Robotic Appearance Have Human-like Preferences? *PloS one* 11, e0146310. doi:10.1371/journal.pone.0146310

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is Psychology Suffering from a Replication Crisis? what Does "Failure to Replicate" Really Mean? *Am. Psychol.* 70, 487–498. doi:10.1037/a0039400

Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers Are Social Actors," in Proceedings of the SIGCHI conference on Human factors in computing systems, Boston, Mass, April 1994, 72–78. doi:10.1145/191666.191703

Nilsson, N. J. (2014). *Understanding Beliefs*. Cambrige: MIT Press.

Norenzayan, A., and Nisbett, R. E. (2000). Culture and Causal Cognition. *Curr. Dir. Psychol. Sci.* 9, 132–135. doi:10.1111/1467-8721.00077

Pak, R., Crumley-Branyon, J. J., de Visser, E. J., and Rovira, E. (2020). Factors that Affect Younger and Older Adults' Causal Attributions of Robot Behaviour. *Ergonomics* 63, 421–439. doi:10.1080/00140139.2020.1734242

Reeves, B., and Nass, C. (1996). *The media Equation: How People Treat Computers, Television, and New media like Real People*. Cambridge, UK: Cambridge University Press.

Riggio, H. R., and Garcia, A. L. (2009). The Power of Situations: Jonestown and the Fundamental Attribution Error. *Teach. Psychol.* 36, 108–112. doi:10.1080/00986280902739636

Robinson, J., and McArthur, L. Z. (1982). Impact of Salient Vocal Qualities on Causal Attribution for a Speaker's Behavior. *J. Personal. Soc. Psychol.* 43, 236–247. doi:10.1037/0022-3514.43.2.236

Ross, L. D., Amabile, T. M., and Steinmetz, J. L. (1977). Social Roles, Social Control, and Biases in Social-Perception Processes. *J. Personal. Soc. Psychol.* 35, 485–494. doi:10.1037/0022-3514.35.7.485

Ross, L. (1977). The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process. *Adv. Exp. Soc. Psychol.* 10, 173–220. Elsevier. doi:10.1016/s0065-2601(08)60357-3

Sætra, H. S. (2021). Robotomorphy. *AI and Ethics*, 1–9. doi:10.1007/s43681-021-00092-x

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., and Sandini, G. (2013). Robots Can Be Perceived as Goal-Oriented Agents. *Is* 14, 329–350. doi:10.1075/is.14.3.02sci

Short, J., Williams, E., and Christie, B. (1976). *The Social Psychology of Telecommunications*. Toronto; London; New York: Wiley.

Snyder, M. L., and Frankel, A. (1976). Observer Bias: A Stringent Test of Behavior Engulfing the Field. *J. Personal. Soc. Psychol.* 34, 857–864. doi:10.1037/0022-3514.34.5.857

Spence, P. R., Westerman, D., Edwards, C., and Edwards, A. (2014). Welcoming Our Robot Overlords: Initial Expectations about Interaction with a Robot. *Commun. Res. Rep.* 31, 272–280. doi:10.1080/08824096.2014.924337

Taylor, S. E., and Fiske, S. T. (1975). Point of View and Perceptions of Causality. *J. Personal. Soc. Psychol.* 32, 439–445. doi:10.1037/h0077095

Thellman, S., Silvervarg, A., and Ziemke, T. (2017). Folk-psychological Interpretation of Human vs. Humanoid Robot Behavior: Exploring the Intentional Stance toward Robots. *Front. Psychol.* 8, 1962. doi:10.3389/fpsyg.2017.01962

Walster, E. (1966). Assignment of Responsibility for an Accident. *J. Personal. Soc. Psychol.* 3, 73–79. doi:10.1037/h0022733

Wortham, R. H., Theodorou, A., and Bryson, J. J. (2017). "Improving Robot Transparency: Real-Time Visualisation of Robot Ai Substantially Improves Understanding in Naive Observers," in 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN), Lisbon, Portugal, August 28–September 1, 2017 (IEEE), 1424–1431. doi:10.1109/roman.2017.8172491

Ziemke, T., Thill, S., and Vernon, D. (2015). "Embodiment Is a Double-Edged Sword in Human-Robot Interaction: Ascribed vs. Intrinsic Intentionality," in Proceedings of the workshop on cognition: A bridge between robotics and interaction, Portland, 9–10.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Robotics and AI

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/frobt.2021.781985

Who Wants to Grant Robots Rights?

Maartje M. A. De Graaf[1]*, Frank A. Hindriks[2] and Koen V. Hindriks[3]

[1]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands, [2]Department of Ethics, Social and Political Philosophy, University of Groningen, Groningen, Netherlands, [3]Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

The robot rights debate has thus far proceeded without any reliable data concerning the public opinion about robots and the rights they should have. We have administered an online survey ($n = 439$) that investigates layman's attitudes toward granting particular rights to robots. Furthermore, we have asked them the reasons for their willingness to grant them those rights. Finally, we have administered general perceptions of robots regarding appearance, capacities, and traits. Results show that rights can be divided in sociopolitical and robot dimensions. Reasons can be distinguished along cognition and compassion dimensions. People generally have a positive view about robot interaction capacities. We found that people are more willing to grant basic robot rights such as access to energy and the right to update to robots than sociopolitical rights such as voting rights and the right to own property. Attitudes toward granting rights to robots depend on the cognitive and affective capacities people believe robots possess or will possess in the future. Our results suggest that the robot rights debate stands to benefit greatly from a common understanding of the capacity potentials of future robots.

Keywords: capacities, reasons, rights, robots, traits

OPEN ACCESS

Edited by:
Anne Gerdes,
University of Southern Denmark,
Denmark

Reviewed by:
Pekka Antero Mäkelä,
University of Helsinki, Finland
Matthew Studley,
University of the West of England,
United Kingdom

***Correspondence:**
Maartje M. A. De Graaf
m.m.a.degraaf@uu.nl

Specialty section:
This article was submitted to
Ethics in Robotics and Artificial
Intelligence,
a section of the journal
Frontiers in Robotics and AI

Received: *23 September 2021*
Accepted: *29 November 2021*
Published: *13 January 2022*

Citation:
De Graaf MMA, Hindriks FA and
Hindriks KV (2022) Who Wants to
Grant Robots Rights?
Front. Robot. AI 8:781985.
doi: 10.3389/frobt.2021.781985

1 INTRODUCTION

Human beings have inalienable rights that are specified in the Universal Declaration of Human Rights. But other entities can have rights too. Animals are commonly taken to have moral rights (Regan, 2004). And organizations have legal rights, including the right to own property and enter into contracts (Ciepley, 2013). But what about robots? Should they have rights? People spontaneously infer intentionality and mind when encountering robots which shows that people cognitively treat robots as social agents (de Graaf and Malle, 2019). But do robots have moral standing, as humans and animals do? Or do they merely have legal rights, just as organizations?

Agents can have moral standing as moral patients. For instance, animals are moral patients because they can suffer. More generally, a moral patient is an agent that can be wronged (Gunkel, 2012). If moral patients have rights, these serve to protect them from such wrongdoings. Agents can also have moral standing as moral agents. Human beings are moral persons, because they are rational and because certain things matter to them. Some of their rights allow or enable them to develop themselves or to live the kind of life they value. The debate about robot rights is commonly framed in terms of moral patiency (Gunkel, 2018). This suggests that they are meant to prevent others from wronging robots.

A third alternative has been proposed by Gunkel (2012), Gunkel (2018) and Coeckelbergh (2010), Coeckelbergh (2021), who defend a social-relational approach to robot rights. Moral patiency and personhood are properties of agents. According to the social-relational approach, the moral standing of robots depends instead on the social relations between humans and robots. Instead of being defined by its attributes, a robot's moral status should be based on people's social responses to robots

(Gunkel, 2018), on how people relate to them, and on the value they have to humans (Coeckelbergh, 2021). In light of this, the social-relational approach can be regarded as human-centered. This is an interesting development particularly because robots cannot suffer and do not value things, which makes it problematic to grant them rights on the basis of their intrinsic properties.

The law treats organizations as legal persons. This notion of legal personhood is often said to be a legal fiction because organizations are not really persons. Because of this legal fiction, they can be granted legal rights. Such rights protect the interests of human beings. Robots might be granted legal rights for the same reason, but this would mean that we have to regard them as legal persons. However, the idea of legal robot rights also has met with controversy.

In 2016, the EU's Committee on Legal Affairs suggested that "the most sophisticated autonomous robots" can have "the status of electronic persons with specific rights and obligations." This committee requested a study on future civil law rules for robotics. This study was commissioned, supervised, and published by the "Policy Department for Citizens' Rights and Constitutional Affairs,"[1] resulting in a resolution by the Parliament.[2] The study aimed to evaluate and analyze a number of future European civil law rules in robotics from a legal and ethical perspective. In an open letter, a coalition of politicians, AI/robotics researchers, industry leaders, health specialists, and law and ethics experts expressed concerns about this.[3] They were worried in particular by the call on the EU commission to explore the implications of creating a specific legal status for robots to address issues related to, for example, any damage robots may cause.

At the same time, others have argued that we need to consider legal personhood for robots because current legal concepts of, for example, responsibility and product liability are no longer sufficient for ensuring justice and protecting those whose interests are at stake (Laukyte, 2019). Thus, robots challenge the law and legal institutions in new ways (Calo, 2015). This is vividly illustrated by the fact that a robot has already been granted citizenship rights (Wootson, 2017).

On the whole, there is little consensus on whether robots should have rights (see Darling (2016), Gunkel (2014), Levy (2009), Schwitzgebel and Garza (2015), Tavani (2018) for some proponents) or not (see Basl (2014), Bryson et al. (2017) for some opponents of this view). Others, such as Gerdes (2016) and Gunkel (2018), have argued that we should at least keep the possibility of granting rights to robots open. These conflicting views raise the question whether and how the debate can progress.

So far, the debate has involved mainly legal experts, philosophers, and policy makers. We, along with Wilkinson et al. (2011), believe that it will be useful to engage the public in the debate about robot rights. Rather than engaging in the debate ourselves, we have conducted an exploratory study investigating people's attitudes toward robot rights through an online survey. To the best of our knowledge, this is the first study that explores layman's opinions on granting robots rights. The main goals are *1*) to examine which reasons people find convincing for granting robot rights and *2*) how willing they are to grant such rights, while *3*) also administering people's general perceptions of robots (appearance, mental capacity, and human-likeness) and *4*) investigating how these relate to their position on robot rights.

Our article is organized as follows. **Section 2** justifies the design of the survey. It embeds it in the literature, it discusses contemporary psychological findings on people's perceptions of robots, and it explains how the rights we consider relate to existing declarations of rights. **Section 3** presents our research design and **section 4** presents our findings. **Section 5** discusses how these results relate to existing findings in HRI research, draws various conclusions, and points to future research directions.

2 THEORETICAL BACKGROUND AND SURVEY DESIGN

Our work empirically investigates people's attitudes toward the issue of granting robots rights by means of an online survey. This section introduces and substantiates the four main survey sections including items on the willingness to grant particular rights to robots in **Section 2.1**, how convincing several reasons are for granting robot rights in general in **Section 2.2**, the belief future robots may one day possess certain capacities and traits in **Section 2.3**, and a general image people have when picturing a robot in **Section 2.4**.

2.1 Rights

The main question that we are interested in here is what everyday people think about the kinds of rights (qualifying) robots deserve. We have broadly surveyed rights that have been granted or proposed for people (human beings), animals, corporations, and, more recently, specifically for robots. As we believe we should at least try to refrain from applying clearly biological categories to robots, we have rephrased our list of rights to match the (apparent) needs of robots, which inherently differ from biological entities (Jaynes, 2020). We have also tried to keep the formulation of rights concrete, simple, and short. As it is not possible to exhaustively determine what the needs (if any) of (future) robots will be, our list may not be complete even though we have tried to compile a list that is as comprehensive as possible. **Table 1** lists the rights used in our study, where the *Source* column indicates the source from which we have derived a right. We refer to rights (and reasons below) by table and row number, for example, 1.1 refers to the right to make decisions for itself. This section discusses how we have translated existing rights to robot rights.

[1] Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics, Committee on Legal Affairs, https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect, accessed February 23, 2020.

[2] European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017IP0051, accessed October 5, 2020.

[3] Open Letter to the European Commission Artificial Intelligence and Robotics, http://www.robotics-openletter.eu/, accessed August 13, 2020.

TABLE 1 | List of robot rights used in the online survey.

Nr	Right	Source
	Should robots have the right to . . .	
1	make decisions for itself	ICESCR Art 1
2	select and block services that it provides	ICESCR Art 6
3	receive fair wages for the work they perform	ICESCR Art 7
4	access energy to recharge themselves	ICESCR Art 11
5	receive updates and maintenance	ICESCR Art 12
6	evolve and develop new capabilities over time	ICESCR Art 13
7	shape and form their own biography	ICCPR Art 6
8	not to be abused either physically or in any other way	ICCPR Art 7
9	be free to leave and return to any country, incl. its own	ICCPR Art 12
10	a fair trial	ICCPR Art 14
11	have freedom of expression through any media of their choice	ICCPR Art 19
12	collectively pursue and protect robot interests	ICCPR Art 22
13	vote for public officials	ICCPR Art 25
14	be elected for political positions	ICCPR Art 25
15	own property	UDHR Art 17
16	the pursuit of happiness	DAW Art 1
17	copy and duplicate themselves	DAW Art 5
18	not to be terminated indefinitely	DAW Art 6
19	enter into contracts	Ciepley, (2013)
20	store and process data they collect	Laukyte, (2019)

2.1.1 Human Rights

Human rights have been documented in the Universal Declaration of Human Rights (UDHR).[4] They have been laid down in two legally binding international agreements, the International Covenant on Civil and Political Rights (ICCPR)[5] and the International Covenant on Economic, Social and Cultural Rights (ICESCR)[6], both adopted in 1966. The rights that feature in these agreements are very different, particularly regarding their means of implementation.

The ICESCR contains economic, social, and cultural rights. These rights were considered to require a proactive role of the state involving financial and material resources. From the ICESCR, we derived rights 1.1-6. For 1.1, we changed "self-determination" into "make decisions for itself" to be more concrete. We assume that robots will be designed to provide specific services to humans (as per the origin of their name, cf., Oxford English Dictionary). As the right to work pertains to "the opportunity to gain his living by work he freely chooses," we reformulated 1.2 in terms of the right to select or block services. As Chopra and White (2004) point out, the ability to control money is important in a legal system since "without this ability a legal system might be reluctant to impose liabilities" on robots; we, therefore, included 1.3. Since robots do not need food (they are artificial physical machines) but do need energy, we have 1.4. We translated "physical and mental health" into "updates and maintenance" (1.5) and "education" into "new capabilities" (1.6).

The ICCPR enumerates a number of civil and political rights or "classic freedom rights." States enforce these rights primarily by not interfering with their citizens. In other words, they are to refrain from action in these fields. From the ICCPR we derived rights 1.7-14. To be suitable for our investigation, we had to adjust them in several respects. To avoid the strong biological connotations of life, we refer to forming a biography in 1.7, in line with Wellman (2018): "A life is a process that involves both goal-directed activities and projects that may succeed or fail and memories of what one has done in the past and what has befallen one [. . .]. The concept of a life is a biographical not a biological concept." We preferred "abuse" over "torture" in 1.8 though we recognize this does not cover "cruel punishment" which may be covered at least in part by 1.18. Right 1.10 was abbreviated to its core. Similarly, we included "freedom of expression" but only in part; we excluded references to (robot) "conscience" and "religion" in 1.11. Furthermore, we translated "freedom of association" and "trade unions" into the collective pursuit and protection of robot interests in 1.12. We split ICCPR Article 25 into two separate rights (as for robots they may have very different consequences, for example, in combination with 1.17). We chose to leave the mechanism of a "secret ballot" implicit. Finally, we derived 1.15 from the UDHR. We believe that most other articles from these declarations and covenants are covered (more or less) already by the rights that we have included or are (clearly) not applicable to robots.

2.1.2 Animal Rights

Rights for nonhuman animals vary greatly by country. Some countries legally recognize nonhuman animal sentience. Others do not even have anti-cruelty laws. We derived three rights from The Declaration on Animal Rights (DAW)[7] that were not yet covered by the rights discussed above. The declaration is still a draft and not yet a law, as most of the human rights are, though animal law exists and is continuously evolving in many countries.

Only the Declaration on Animal Rights refers explicitly to "the pursuit of happiness" as a right, which is why we included 1.16 as a separate item. To avoid the perhaps strong biological connotations with "reproduce" and "offspring", we translated these into "copy and duplicate" in 1.17, which we believe is the more appropriate analogical terminology for robots. Similarly, we translated, for example, "slaughtered" and "killed" to "terminated indefinitely" in 1.18. We have added the qualification "indefinitely" to meet the objection of Jaynes (2020), who argues that "depriving power to the [robot] cannot be considered an act of murder, as the [robot]'s "personality" will resume once power has been restored to the system." Finally, there might be a relation between this right and the right to life. After all, terminating a robot indefinitely would make shaping its own biography impossible. Even so, some argue that only those that have the potential for self-determination (ICCPR Article 1)

[4] Universal Declaration of Human Rights, https://www.un.org/en/universal-declaration-human-rights/, accessed on March 1, 2020, which was adopted in 1948 by the United Nations General Assembly.

[5] https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx, accessed March 1, 2020.

[6] https://www.ohchr.org/en/professionalinterest/pages/cescr.aspx, accessed March 1, 2020.

[7] https://declarationofar.org/, accessed March 1, 2020.

and moral action (autonomy) can have a right to life. We regard the two as sufficiently distinct to include both.

2.1.3 Corporate Rights

Corporations are created by means of a corporate charter, which is granted by the government. They receive their rights from their charter (Ciepley, 2013). As mentioned in the introduction, corporations are often seen as legal fictions. Chief Justice Marshall puts it in Dartmouth as follows: “A corporation is an artificial being, invisible, intangible, and existing only in contemplation of law. Being the mere creature of law, it possesses only those properties which the charter of its creation confers upon it” (Dartmouth College v. Woodward 1819, 636; our emphasis). Perhaps the most important right that corporations have is the right to enter into contracts (Ciepley, 2013). As it seems possible for robots to possess it, we include it as right 1.19.

2.1.4 Robot-specific Rights

Finally, inspired by Laukyte (2019), we add right 1.20 to store and process data which arguably is associated specifically with robots.

TABLE 2 | List of reasons used in the online survey.

Nr	Reason
	How convincing is it to grant robots rights when . . .
1	they can perceive the world around them
2	they can experience pain
3	they can experience pleasure
4	they can have feelings
5	when they can pay attention
6	when they have preferences
7	they can have memories
8	they can use language
9	they can independently make decisions and act on their own
10	they can take their own moral considerations into account
11	they have a conscience
12	they can make rational decisions
13	they are super-intelligent
14	human beings can no longer be held responsible for what robots do
15	they can learn
16	they appear humanlike
17	they can move around
18	they can understand others
19	they have a unique personality
20	they can love people
21	it is convenient to do so

2.2 Reasons for Granting Robots Rights

Many (combinations of) reasons have been put forward for granting robots rights. Miller (2015) maintains that robots “with capacity for human-level sentience, consciousness, and intelligence” should be considered entities that “warrant the same rights as those of biological humans.” Tavani (2018) thinks that a robot should have consciousness, intentionality, rationality, personhood, autonomy, and sentience to be eligible for rights. Strikingly, many of these properties are requirements for moral personhood. Laukyte (2019) states that the increasing autonomy, intelligence, perceptiveness, and empathy of robots shift our view away from robots as mere tools. These are among the main reasons for granting robots rights. Based on a review of the literature, we have tried to identify the main reasons that have been discussed so far (see **Table 2**).

2.2.1 Consciousness

Consciousness is an important reason in the literature for granting robots rights. Levy (2009) claims that robots should be treated ethically by “virtue of their exhibiting consciousness.” It is common to distinguish between two kinds of consciousness, phenomenal consciousness on the one hand and access or functional consciousness on the other (Block, 1995; Torrance, 2012). Phenomenal consciousness requires sentience. As such, it is experiential and subjective. Think, for instance, of seeing, hearing, smelling, tasting, and feeling pain. Phenomenal conscious states encompass sensations, perceptions, feelings, and emotions. In contrast, access consciousness concerns awareness and plays an essential role in reasoning (Block, 1995). It is representational and makes mental content available for evaluation, choice behavior, verbal report, and storage in working memory (Colagrosso and Mozer, 2005).

Torrance (2012) states that “it is the phenomenal features of consciousness rather than the functional ones that matter ethically.” The main related reason that is often cited for granting entities moral status and rights is that they can suffer: they can experience pain from physical or emotional harm. The ability to (physically) suffer has also been one of the main reasons for granting rights to animals (Singer, 1974). We include the concrete reason items 2.1-5 for perception, suffering, experiencing pleasure, feelings, and attention. Note, however, that it is contested whether robots will ever be able to feel pain (see Levy (2009) contra versus Kuehn and Haddadin (2017) pro). We did not add a separate item for “consciousness.” Given how complex the notion is, this would not be meaningful.

Insofar as access consciousness is concerned, Freitas (1985) argues that “any self-aware robot that speaks [a language] and is able to recognize moral alternatives” should be considered a “robot person.” The EU draft report mentioned in the introduction also refers to the ability of robots to “make smart autonomous decisions or otherwise interact with third parties independently” to grant robots the status of an electronic personality. These items correspond to cognitive skills that humans have. We include reason items 2.6-9 for access-related phenomena. Although decision making involves preferences, we regard it as important to add it as a separate item.

2.2.2 Autonomy

Another reason for assigning rights has been the ability to make decisions and perform actions independently, without any human intervention. This capability corresponds to the cognitive ability of humans to make decisions. It is not sufficient that a system can act without human intervention. That would be mere automation (the machine can act automatically) and does not capture the richer sense of what autonomy is. “To be autonomous, a system must have the capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation."[8] Tessier (2017), moreover, adds that such decision making should be based on an understanding of the current situation.

Independent decision making and acting (without human intervention) is only one aspect of the notion of autonomy. Another reason for assigning rights is the ability to make decisions and to live your life according to your own moral convictions. Borenstein and Arkin (2016) also note that there is a difference in how the term "autonomy" is normally used in ethics in contrast with how it is used within AI: "the term 'autonomy' in the sense of how it is normally defined within the realm of ethics (i.e., having the meaningful ability to make choices about one's life); within the realm of robotics, 'autonomy' typically refers to a robot or other intelligent system making a decision without a 'human in the loop.'" The ability to distinguish right from wrong also has been put forward as an argument in favor of legal personhood (Chopra and White, 2004). This discussion motivated items 2.10-11.

2.2.3 Rationality and Super-Intelligence

Rationality has been put forward as an important reason why humans have moral standing. According to Nadeau, "only machines can be fully rational; and if rationality is the basic requirement for moral decision making, then only a machine could ever be considered a legitimate moral agent. For Nadeau, the main issue is not whether and on what grounds machines might be admitted to the population of moral persons, but whether human beings qualify in the first place" (Gunkel (2012); see also Sullins (2010)). Solum (1992) argues that intelligence is a criterion for granting rights. Robots may become much smarter than the best human brains in practically every field. When robots outperform humans on every cognitive or intellectual task and become super-intelligent, some argue we should assign them robot rights. This discussion motivated items 2.12-13.

2.2.4 Responsibility Gaps

In a communication to the members of the EU Parliament, before they voted on the Resolution on Civil Law Rules of Robotics on February 16, 2017, the intention to grant a legal status to robots was clarified as follows: "In the long run, determining responsibility in case of an accident will probably become increasingly complex as the most sophisticated autonomous and self-learning robots will be able to take decisions which cannot be traced back to a human agent." Another argument that has been put forward is that if robots are able to perform tasks independently without human intervention, it will be increasingly difficult to point responsibility to a specific person or organization when something goes wrong (Danaher, 2016). Some scholars therefore propose that moral and legal responsibility should at some point be extended to robots (Wiener, 1954). This motivates reason 2.14. We added 2.15 because the ability of robots to learn has also been cited as a key reason for responsibility gaps, e.g., Matthias (2004).

2.2.5 Humanlike Appearance and Embodiment

The fact that robots will at some point become indistinguishable from humans, both in their looks and the ways they behave, is for some scholars a reason to assign rights to robots. If robot appearance becomes very similar to that of human beings, one could argue that the basis for making a moral distinction between robots and humans is no longer tenable (Darling, 2016; Gunkel, 2018). This motivated item 2.16. Item 2.17 has been added to also emphasize the embodiment of robots and their physical ability of moving on their own capacity, as perhaps having the looks without being able to move will not do.

2.2.6 Mind Perception, Personality, and Love

Understanding others' minds (Gray et al., 2007; Gray et al., 2012) also seems relevant as Laukyte (2019) states that empathy of robots shifts our view away from robots as mere tools, and, moreover, this capacity matches with an item in the mental capacity scale (Malle, 2019). The notion of understanding others also raises the question about one's own unique personality or identity and related notions of connectedness such as love as reasons for having rights, which motivated introducing items 2.18-20.

2.2.7 Convenience

Finally, item 2.21 was added because one could also argue that from a more pragmatic stance, we should grant robots rights "simply" because they play a significant role in our society and granting robots rights may depend on "the actual social necessity in a certain legal and social order" (van den Hoven van Genderen, 2018).

2.3 Psychological Factors

People's willingness to grant robot rights could result from their perceptions of future robots, and could be linked to the conceptions of moral patiency (and agency) presented in **Section 1** by linking the philosophical interpretations of a robot's moral standing to foundations in moral psychology research. Balancing on the intersection of philosophy and psychology, moral psychology research revolves around moral identity development and encompasses the study of moral judgment, moral reasoning, moral character, and many related subjects at the intersection of philosophy and psychology. Questions on how people perceive an entity's moral status is often investigated with theories of mind perception.

Effects of human-likeness in human–robot interaction have been profoundly discussed (Fink, 2012; Złotowski et al., 2015). In our survey, we aimed to go beyond a robot's anthropomorphic form to focus on the potential humanness of robots. A research body on humanness has revealed specific characteristics perceived as critical for the perception of others as human and distinguishes two senses of humanness (Haslam, 2006), which we included in our survey. First, *uniquely human* characteristics define the boundary that separates humans from the related

[8] Defense Science Board Summer study on autonomy, United States Defense Science Board, https://www.hsdl.org/?view&did=794 641, accessed March 13, 2020.

category of animals and includes components of intelligence, intentionality, secondary emotions, and morality. Denying others such characteristics is called *animalistic dehumanization* in which others are perceived as coarse, uncultured, lacking self-control, and unintelligent, and their behaviors are seen as driven by motives, appetites, and instincts. Second, *human nature* characteristics define the boundary that separates humans from nonliving objects and includes components of primary emotions, sociability, and warmth. Denying others such characteristics is called *mechanistic dehumanization* in which others are perceived as inert, cold, and rigid, and their behavior is perceived as caused rather than propelled by personal will.

These two senses of humanness can also be linked to the perception of mind. According to Gray et al. (2007), the way people perceive mind in other human and nonhuman agents can be explained by two factors: agency and experience, where agency represents traits such as morality, memory, planning, and communication, and experience represents traits such as feeling fear, pleasure, and having desires. The agency dimension of mind perception corresponds to uniquely human characteristics, and the experience dimension links to human nature characteristics (Haslam et al., 2012). These two dimensions are linked to perceptions of morality such that entities high in experience and entities high in agency are considered to possess high moral agency (Gray et al., 2007) and thus deserving of (moral) rights.

However, perceiving mind, and consequently deserving of morality (Gray et al., 2007) and presumably rights, is regarded as a subtle process (de Graaf and Malle, 2019). In particular, the dual-dimensional space of mind perception has been challenged as several studies failed to replicate especially the agency dimension, e.g., Weisman et al. (2017). A recent series of studies provides consistent evidence that people perceive mind on three to five dimensions (i.e., positive and negative affect, moral and mental regulation, and reality interaction) depending on an individual's attitude toward the agent (e.g., friend or foe) or the purpose of mind attribution (e.g., interaction or evaluation) (Malle, 2019), and our survey has therefore administered the mental capacity scale of Malle (2019).

In summary, previous HRI research shows that people's ascription of humanness as well as mind capacity to robots affects how people perceive and respond to such systems. In line with the social-relational perspective to a robot's moral standing (Gunkel, 2012; Gunkel, 2018; Coeckelbergh, 2010; Coeckelbergh, 2021), we will investigate how such perceptions of humanness and mind influence people's willingness to granting rights to robots.

2.4 Appearance of Robots

Although what constitutes a robot can significantly vary between people (Billing et al., 2019), most people, by default, appear to have a humanlike visualization of a robot (De Graaf and Allouch, 2016; Phillips et al., 2017). Nevertheless, what appearance people have in mind is relevant for answering the question whether they are eligible for rights. It is not clear up front which kind of robots (if any) deserve rights (Tavani, 2018). Here, we only assume that robots are artificial (i.e., not natural, nonbiological) physically embodied machines. To get a basic idea of people's perception of what a robot looks like, we include a simple picture-based robot scale (Malle and Thapa, 2017), **Figure 1** in our survey.

3 METHODS

To examine layman's opinions regarding robot rights, we have conducted an online survey administering participants' willingness to grant particular rights to robots and their indication of how convincing several reasons are to grant those rights, while also administering people's general perceptions of robots.

3.1 Procedure and Survey Design

After participants gave their consent, we introduced the survey topic describing that "[technological advancements], amongst other things, has initiated debates about giving robots some rights" and that "we would like to learn about [their] own opinions on several issues regarding the assignment of rights to robots." The survey consisted of four randomly shown blocks (see **Section 2**) to avoid any order effects. The survey ended with questions regarding basic demographics, professional background, and knowledge and experience with robots. Average completion time of the survey was 11 (SD = 4:18) minutes, and participants' contribution was compensated with $2.

The first block of the online survey contained one question asking participants which kind of robot appearance (see **Figure 1**) best resembles their image of a robot in general. The second and third block contained the reasons and rights items, respectively, of which the item selection was discussed in **Section 2**. The structure of each of the reason items was as follows and had the same format: "Suppose that robots [features]. How convincing do you think it is to grant rights to robots*. . .when [reason]*." The [feature] slot is filled with capacities or features that robots will eventually possess to frame the question and put participants in a state of mind where they would presume these to be the case for (future) robots. The [reason] slot is filled with one of the 21 reasons from **Table 2**. For example, the item for the first reason is: "Suppose that robots can see, hear, smell, and taste. How convincing do you think it is to grant rights to robots*. . .when they can perceive the world around them*." Participants were instructed to rate how appropriate they thought it would be to grant rights on a 7-points Likert scale. The format for the rights items is "Robots should have the right to [right]" where the [right] slot is filled with one of the rights from **Table 1**. For example, the item for the first right is: "Robots should have the right to*. . .make decisions for themselves*." and participants were asked to rate how strongly they would oppose or favor granting the right on a 7-point Likert scale. The fourth block administered participants' perceptions of future robots. To measure perceptions of capacities, we used the mental capacity scale developed by Malle (2019) consisting of the subscales affect (α = 0.94), cognition (α = 0.90), and reality interaction (α = 0.82). To measure perceptions of traits, we used the dehumanization scale developed by Haslam (2006) consisting of the subscales uniquely human (α = 0.85) and human nature (α = 0.98).

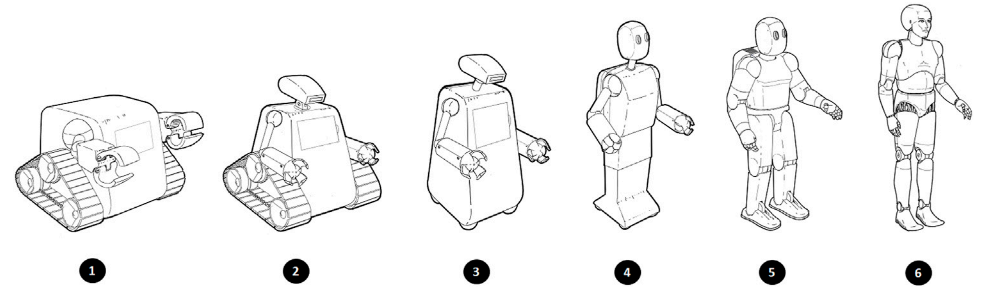

FIGURE 1 | Robot appearance scale.

3.2 Participants

In April 2020, we initially recruited 200 USA-based participants from Amazon Mechanical Turk (De Graaf et al., 2021). In May 2021, we replicated our study by recruiting 172 EU-based participants from Amazon Mechanical Turk and 200 participants from Asia using Prolific. All participants, from either platform, had an approval rate of >95%. For the EU and Asia samples, we administered a Cloze Test (Taylor, 1953) to ensure a good command in English, which led to the exclusion of 72 participants from Europe and 19 participants from Asia. In addition, 39 participants from the Asia sample were removed from further analysis because they had indicated growing up in Europe or the USA. The final data set used in our analyses included $n = 439$ participants (USA: $n = 200$, EU: $n = 97$, Asia: $n = 142$). In the EU sample, most participants were living in Italy ($n = 36$), Spain ($n = 25$) or Germany ($n = 17$). In the Asia sample, most participants were born and raised in China ($n = 73$), South Korea ($n = 34$), or Singapore ($n = 17$).

The complete sample included 53.3% men, 46.0% women, and 0.7% identified as gender-nonbinary. Participants' age ranged from 20 to 71 ($M = 35.5$, SD = 11.2), their educational level ranged from high school degree (23.2%) and associates degrees (11.4%) to bachelor's, master's, and doctoral degrees (65.1%), and 23.5% had a profession in computing and engineering. Most participants indicated having no or little knowledge about robots (52.1%) and never or rarely encounter robots in their daily life (71.9%), and participants mainly hold humanoid images of robots (61.3% selected picture five or six on the robot appearance scale). Measures on the robot appearance scale correlated only with the interaction capacity scale—and did so weakly ($r = 0.181$, $p = 0.01$)—and was therefore excluded from further analysis.

4 RESULTS

4.1 Factor Analysis

As a first step, we conducted two separate factor analyses to reduce the individual items into a fewer number of underlying dimensions that characterize: *1*) the types of rights people are willing to assign to robots; and *2*) the types of reasons they consider for doing so. There were no outliers (i.e., Z-score of >3.29). Both sets of items were independently examined on several criteria for the factorability of a correlation. First, we observed that all 20 rights and all 21 reasons correlated at least 0.3 with at least one other right or reason, respectively, suggesting reasonable factorability. Second, the Kaiser-Meyer-Olkin measure of sampling adequacy was 0.97 for rights and 0.96 for reasons, well above the commonly recommended value of 0.6. Bartlett's test of sphericity was significant in both sets, for rights ($\chi^2(190) = 6518.97, p < 0.001$) and for reasons ($\chi^2(210) = 6822.39$, $p < 0.001$), respectively. The diagonals of the anti-image correlation matrix were also all over 0.5. Finally, the communalities were all above 0.35, further confirming common variance between items. These overall indicators deemed factor analysis to be appropriate.

An eigenvalue Monte Carlo simulation (i.e., a parallel analysis) using the method described in (O'connor, 2000) indicated the existence of two and potentially three underlying dimensions for both the reasons and rights items. Solutions for both two and three factors were explored. We executed the factor analysis using an Alpha factors extraction (a method less sensitive to non-normality in the data (Zygmont and Smith, 2014)) with Oblimin rotations (allowing correlations among the factors)). A two-factor solution was preferred for both the reason and right items because of *1*) the leveling off of Eigenvalues on the screen plot after two factors; *2*) a low level of explained variance (<4%) of the third factor in both cases; and *3*) the lower number of cross-loading items.

The two reason factors had a total explained variance of 64.3%. Factor 1 revealed ten *cognition reasons* and factor 2 revealed nine *compassion reasons* both with strong factor loadings (>.5; see **Table 3** for the specific items). A total of two items were eliminated because they did not contribute to a simple factor structure and failed to meet a minimum criterion of having a primary factor loading of >.5 and/or had cross-loading of >.4 (i.e., having preferences, and making rational decisions). Internal consistency for each of the sub-scales was examined using Cronbach's alpha, which were 0.93 for both cognition and compassion reasons. No increases in alpha for any of the scales could have been achieved by eliminating more items.

The two rights factors had a total explained variance of 64.1%. Factor 1 revealed thirteen *sociopolitical rights* and factor 2 revealed six *robot rights* both with strong factor loadings (>.5; see **Table 4** for specific items). One item was eliminated because it did not contribute to a simple factor structure and failed to meet a

TABLE 3 | Loading matrix of factor analysis on 21 reasons.

Reasons		Factor 1	Factor 2
		Cognition	**Compassion**
17	Moving around	0.926	–0.238
8	Using language	0.912	–0.112
5	Paying attention	0.852	–0.040
15	Learning	0.717	0.177
16	Appearing humanlike	0.653	0.038
7	Having memories	0.652	0.203
13	Super-intelligence	0.648	0.204
21	Convenience	0.616	–0.025
1	Perceiving the world	0.570	0.312
18	Understanding others	0.537	0.383
6	Having preferences	0.467	0.430
12	Making rational decisions	0.429	0.397
4	Having feelings	–0.186	0.967
11	Having a conscience	–0.146	0.907
10	Moral considerations	0.053	0.821
2	Experiencing pain	–0.057	0.821
20	Loving people	0.119	0.731
3	Experiencing pleasure	0.141	0.681
9	Acting on its own	0.171	0.659
14	Human responsibility impossible	0.128	0.542
19	Having a unique personality	0.377	0.502
	Eigenvalue	10.78	2.73
	% Explained variance	51.3	13.0
	Subscale Cronbach's α	0.93	0.93

TABLE 4 | Loading matrix of factor analysis on 20 rights.

Nr	Right	Factor 1	Factor 2
		Sociopolitical	**Robot**
13	Vote	0.985	–0.229
14	Be elected	0.936	–0.228
15	Own property	0.875	–0.020
17	Duplicate	0.642	–0.007
9	Cross nation borders	0.635	0.217
1	Self-decide	0.598	0.278
3	Fair wages	0.586	0.282
12	Pursue and protect interests	0.570	0.358
18	Not be terminated	0.564	0.250
19	Enter into contracts	0.561	0.261
7	Form own biography	0.560	0.290
2	Block services	0.531	0.276
11	Freedom of expression	0.519	0.389
16	Pursuit of happiness	0.474	0.456
5	Updates and maintenance	–0.138	0.871
4	Access to energy	–0.004	0.765
8	Not be abused	0.077	0.713
6	Self-development	0.209	0.605
10	A fair trial	0.357	0.549
20	Process collected data	0.151	0.504
	Eigenvalue	11.08	1.75
	% Explained variance	55.4	8.7
	Subscale Cronbach's α	0.95	0.88

minimum criterion of having a primary factor loading of $>.5$ and/or had cross-loading of $>.4$ (i.e., pursuit of happiness). Internal consistency for each of the sub-scales was examined using Cronbach's alpha, which were 0.95 for sociopolitical rights and 0.88 for robot rights, respectively. No increases in alpha for any of the scales could have been achieved by eliminating more items.

4.2 Cluster Analysis

As a second step, we explored the data using cluster analysis to classify different groups of people based on their opinions about rights for robots and reasons to grant those. A hierarchical agglomerate cluster analysis was performed using Ward's method as a criterion for clustering (Ward, 1963; Murtagh and Legendre, 2011). Clusters were initially considered by visually analyzing the dendrogram (Bratchell, 1989) while considering the iteration history, significance of the F statistics, and the number of individuals in each cluster. This was done to ensure the cluster solution was stable, that there was a clear difference between clusters, and that each cluster was well represented ($n > 15\%$).

The analysis resulted in three clearly distinguishable clusters. Chi-square tests revealed significant demographic differences between the clusters in terms of age ($\chi^2(4) = 10.78$, $p = 0.029$) and continent ($\chi^2(3) = 25.54$, $p < 0.001$), and marginally significant differences for educational level ($\chi^2(4) = 7.86$, $p = 0.097$) and robot encounters ($\chi^2(2) = 5.28$, $p = 0.071$). No significant differences were found for gender ($\chi^2(2) = 0.12$, $p = 0.941$), profession ($\chi^2(2) = 0.22$, $p = 0.896$), or robot knowledge ($\chi^2(2) = 3.97$, $p = 0.138$). Participants in cluster 1 ($n = 99$) are more likely people from the US ($z = 2.9$) and possibly not aged 55 and older ($z = -1.2$), have a lower educational level ($z = 1.9$), and encounter robots occasionally or frequently ($z = 1.8$). Participants in cluster 2 ($n = 245$) are more likely people from Asia ($z = 2.5$) and possibly aged 30 and younger ($z = 1.4$), and possibly have a higher educational level ($z = 2.1$). Participants in cluster 3 ($n = 93$) are more likely people from Europe ($z = 1.9$) and aged 55 and older ($z = 2.7$), and possibly have never or rarely encountered robots ($z = 1.9$).

A series of one-way ANOVA tests showed significant differences between the three clusters in assessments of robot capabilities and traits as well as their opinions about rights for robots and reasons to grant those. Given a violation of the homogeneity of variance assumption and the unequal sample sizes between the three clusters, we have reported the Welch's F-statistics (Tomarken and Serlin, 1986) (see **Table 5**). These combined results indicate that participants in cluster 1 seem to hold a *cognitive affective view* on robots being more positive toward granting robots rights, deeming the reasons for granting rights to be more convincing, and believing in higher potentials of future robot capacities and traits. Participants in cluster 2 seem to hold a *cognitive but open-minded view* on robots being more positive toward granting rights to robots as well as the cognitive and interaction capacities of robots, but being more skeptical toward the affective capacities of future robots while indicating compassion reasons to be convincing for granting robots rights. Participants in cluster 3 seem to hold a *mechanical view* on robots being only positive about future robots' capacity for interaction but being rather negative toward granting rights, nor deeming the

TABLE 5 | Average construct ratings for all participants and per cluster.

	All		Cluster 1		Cluster 2		Cluster 3		Welch's ANOVA		
Construct	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**F(2,434)**	**p**	**Cohen's d**
Capacity											
Cognition	4.02	1.40	5.12	1.10	3.93	1.22	3.10	1.38	69.12	0.000	1.44
Affect	2.67	1.49	3.97	1.60	2.53	1.22	1.65	0.96	75.97	0.000	1.56
Interaction	5.77	1.26	6.35	0.79	5.75	1.14	5.23	1.67	24.87	0.000	0.88
Trait											
Human nature	3.43	1.26	4.53	1.12	3.33	1.03	2.50	1.07	83.01	0.000	2.53
Uniquely human	4.14	1.19	4.97	0.99	4.05	1.02	3.48	1.30	46.35	0.000	1.62
Reason											
Cognition	3.73	1.48	5.15	1.07	3.84	1.07	1.94	0.81	304.77	0.000	2.17
Compassion	4.62	1.74	5.96	0.62	4.90	0.81	2.47	1.16	340.12	0.000	2.37
Right											
Robot	5.02	1.37	6.21	0.58	5.19	0.84	3.32	1.46	1964.96	0.000	2.11
Sociopolitical	3.47	1.46	5.31	0.84	3.34	0.93	1.84	0.72	475.83	0.000	2.40

Tukey HSD significance are at $p < 0.01$ between all pairs.

reasons for granting rights to be convincing, and being generally skeptical about the potentials of future robot capacities and traits.

4.3 Regression Analysis

Given our aim to uncover the minimum number of predictors which significantly explains the greatest amount of variance for both sociopolitical and robot rights, we ran a series of step-wise multiple regressions for each cluster separately.

4.3.1 Explaining Sociopolitical Rights

For cluster 1 (*people with a cognitive affective view on robots*), the capacities, traits, and reasons to assign rights were significant predictors of participants' readiness to grant robots sociopolitical rights ($F(2, 96) = 14.36$, $p < 0.001$). Together, the capacity of cognition ($\beta = 0.420$, $p < 0.001$) and cognition reason ($\beta = - 0.188$, $p = 0.040$) explained 23% of the variance. Readiness to grant sociopolitical rights was for cluster 1 participants associated with beliefs that robots will (eventually) possess cognitive capacities while considering cognition reasons had a negative effect on their readiness to grant sociopolitical rights. For cluster 2 (*people with a cognitive but open-minded view on robots*), the capacities, traits, and reasons to assign rights were significant predictors of participants' readiness to grant robots sociopolitical rights ($F(1, 243) = 57.29$, $p < 0.001$). The capacity of affect ($\beta = 0.437$, $p < 0.001$) was the sole predictor explaining 19% of the variance. Readiness to grant robots sociopolitical rights was for cluster 2 participants associated with beliefs that robots will (eventually) possess affective capacities. For cluster 3 (*people with a mechanical view on robots*), the capacities, traits, and reasons to assign rights were significant predictors of participants' readiness to grant robots sociopolitical rights ($F(3, 87) = 21.94$, $p < 0.001$). Together, the capacity of cognition ($\beta = 0.537$, $p < 0.001$), the trait of uniquely human ($\beta = - 0.246$, $p = 0.028$), and cognition reason ($\beta = 0.421$, $p < 0.001$) explained 41% of the variance. Readiness to grant robots sociopolitical rights was for cluster 3 participants associated with beliefs that robots will (eventually) possess cognition capacities but lacking traits of intelligence, intentionality, secondary emotions, and morality (uniquely human) while considering cognition reasons positively affected their readiness to grant sociopolitical rights.

4.3.2 Explaining Robot Rights

For cluster 1 (*people with a cognitive affective view on robots*), the capacities, traits, and reasons to assign rights were significant predictors of participants' readiness to grant robots rights ($F(1, 97) = 15.09$, $p < 0.001$). The capacity of interaction ($\beta = 0.367$, $p < 0.001$) was the sole predictor explaining 14% of the variance. So, for cluster 1 participants, their belief that robots will (eventually) possess interaction capacities seems to be enough to grant the rights in our robot rights dimension to robots. For cluster 2 (*people with a cognitive but open-minded view on robots*), the capacities, traits, and reasons to assign rights were significant predictors of participants' readiness to grant robots the rights in our robot rights dimension ($F(3, 241) = 17.26$, $p < 0.001$). Together, the capacity of interaction ($\beta = 0.278$, $p < 0.001$), the trait of human nature ($\beta = 0.151$, $p = 0.013$), and compassion reason ($\beta = 0.200$, $p = 0.001$) explained 17% of the variance. So, for cluster 2 participants, besides (eventually) possessing interaction capacities, robots will (eventually) have the traits of primary emotions, sociability, and warmth (human nature) to grant robot rights while considering compassion reasons further positively affected their readiness to do so. For cluster 3 (*people with a mechanical view on robots*), the capacities, traits, and reasons to assign rights were significant predictors of participants' readiness to grant robots the rights in the robot rights dimension ($F(3, 87) = 11.14$, $p < 0.001$). Together, the capacity of cognition ($\beta = 0.304$, $p = 0.002$) as well as cognition ($\beta = 0.209$, $p = 0.045$) and compassion ($\beta = 0.222$, $p = 0.028$) reasons explained 25% of the variance. So, for cluster 3 participants, their readiness to assign the rights in the robot rights dimension to robots was justified by their beliefs that robots will (eventually) possess cognitive capacities while considering both cognition and compassion reasons positively affected their readiness to do so.

5 DISCUSSION

Current discussion on robot rights is dominated by legal experts, philosophers, and policy makers. To consider the opinion of lay persons in the policy debate, in line with the social-relational perspective to a robot's moral standing (Gunkel, 2012, 2018; Coeckelbergh, 2010, 2021), we explored people's attitudes toward the issue of granting rights to robots in an online survey. A factor analysis has again identified two main dimensions for both reasons and rights, replicating our previous findings with the US-only sample (De Graaf et al., 2021). The reason dimensions consist, on the one hand, of mainly *cognition reasons* (e.g., moving around, language, attention, learning) with only two other at face value unrelated items (i.e, humanlike appearance and convenience) as reasons for granting robots rights, and affect-related *compassion reasons* (e.g, feelings, conscience, pain, moral considerations) on the other hand with only one at face value unrelated item (i.e., acting on one's own). It thus appears that people's perspective on robot affect and cognition plays an important role in the context of granting robots rights, which is also in line with the results of our cluster and regression analysis.

The first rights dimension, labeled *sociopolitical rights*, consists mainly of items associated with the freedom to do what one wants (e.g., vote, duplicate, cross borders, self-decide, shape one's biography) and to be treated fairly (e.g., be eligible for election, own property, fair wages). A clearly different second dimension, labeled *robot rights*, mainly consists of items associated with a robot's technical needs to function properly (updates, energy, self-development, process data) and the item to not be abused. One explanation why this last item is also associated with this dimension is that the right to not be abused was perceived as damaging other people's property. These two rights dimensions reveal that people tend to differentiate between more general sociopolitical rights and those associated with a robot's functional needs.

The average ratings for the various scales used in our study show that only the capacity of reality interaction (e.g., learning, verbally communicating, moving, perceiving the world) had high overall agreement that robots can do this well (see **Table 5**). People, thus, generally tend to have a rather positive view on the capabilities of (future) robots regarding their ability to (socially) interact with their environment, irrespective of their user characteristics (e.g., age, gender, continent, robot experience). The interaction capacity also predicts readiness to grant robot rights. The high averages on this scale indicate a high willingness to grant robot rights to robots (except for *people from EU, those aged 55 and older, and those less familiar with robots*, who tend to be more skeptical). Most people (about 80%) thus agree that robots should be updated, have access to energy, process collected data, and not be abused.

This is different for sociopolitical rights (e.g, voting, fair wages, and the right not to be terminated) which *people from cluster 1* (i.e., those who are most likely from the US, and possibly not aged 55 and older, have a lower educational level, and have encountered robots occasionally or frequently) seem to be most willing to grant to robots. This may be explained by our finding that these people are also more optimistic about the possibility that future robots can have affect, cognition, and human traits. Moreover, there is a strict order where people from cluster 1 are significantly more willing to grant sociopolitical rights than people from cluster 2 (i.e., those who are more likely from Asia, and possibly aged 30 and younger and have a higher educational level) followed by people from cluster 3 (i.e., those who are most likely from Europe and aged 55 and older, and possibly have never or rarely encountered robots) being least willing to do so.

Our findings suggest that it is more likely that people from the US are very optimistic about the potential of robots in general and are more likely to assign them rights, people from Asia are positioned somewhere in the middle on these issues, and people from Europe are overall much more skeptical. Our findings are somewhat similar to those of Bartneck et al. (2007) who also find that people from the US are the most positive, more so than Japanese, who appear in turn more positive than Europeans. Although one might be tempted to conclude there is a cultural link between assigning rights to robots from this, more evidence is needed to conclude such a relation. Note that our continent-based samples do *not* match with clusters (sizes differ with the US sample a size of $N = 200$ vs. cluster 1 a size of $N = 99$, the Asia sample a size of $N = 142$ vs. cluster 2 with a size of $N = 245$, and the EU sample with a size of $N = 97$ vs. cluster 3 with a size of $N = 93$). MacDorman et al. (2008) also do not find any evidence for strong cultural differences between the US and Japan. A cultural interpretation of our findings therefore seems premature and would require more research to support such conclusions.

Based on our cluster analysis, we can conclude that *people from cluster 3* (i.e., those who are more likely from Europe and aged 55 and older, and possibly have never or rarely encountered robots) generally have a more *mechanical view* of robots and are more skeptical about robots having cognitive or affective capacities or humanness traits. This is in line with a tendency for mechanistic dehumanization in this group. Because cognition and affect-related reasons are a predictor for this group, only if these capacities will be realized are they willing to grant sociopolitical rights. *People from cluster 2* (i.e., those more likely from Asia, possibly aged 30 and younger, and possibly with a higher education level) have a significantly more positive view and believe robots will have cognitive capacities and human traits, but they are less inclined to believe that robots will have affects, which for them is important to grant sociopolitical rights. This group appears to have a *cognitive view* of robots but is more skeptical about affective capacities. Note that all groups more strongly believe that robots will have cognitive rather than affective capacities (*see* **Table 5**). In contrast, *people from cluster 3* (i.e., those more likely from the US, and possibly not aged 55 and older, have a lower education level, and have encountered robots occasionally or frequently) have a very positive view on all capacities and traits of future robots. It appears that they have a *cognitive-affective view* of robots.

In our analysis, we did not find many strong relations between demographical factors and people's views on assigning them rights (with the exception of age and continent), which is in line with the findings reported in MacDorman et al. (2008) which also does not find such relations. Flandorfer (2012) has reported on a link between age, experience, and attitude toward robots. In this work, it appears a younger age is associated with higher exposure to and more positive views on new technology in general, but we did not find such a trend. Finally, our findings overall are similar to those reported in our previous work (De Graaf et al., 2021) which only analyzed the US sample. One noticeable difference is that in our current analysis, we found only three instead of four clusters which are correlated with the continents associated with the three samples we collected. The fact we had four groups in our previous work is explained by the differences in experience with robots that does not play a differentiating role in our current analysis.

5.1 Limitations and Future Work

As any study, ours has some limitations. First, the three samples from the US, EU, and Asia varied significantly in division of age category and educational level. Regarding age, the US sample had an overrepresentation of people aged 50 and over, and the Asia sample had an overrepresentation of people aged 30 and younger. These demographics are actually quite similar to the actual population demographics in these continents.[9] Regarding educational level, the US sample had an overrepresentation of people with a high school degree, and the Asia sample had an overrepresentation of people with a bachelor's, master's, or doctoral degrees.

Second, participants may have interpreted the survey items differently, particularly the reason items because of their conditional nature. We asked to suppose robots had certain capabilities or features and assess their willingness to grant rights *if* that were the case. Similarly, for the robot rights, which may have been granted more easily because participants read those more as operational requirements for robots rather than as rights. Future work should address any potential difficulty with interpreting these conditionals (Skovgaard-Olsen et al., 2016) to further validate our items and underlying dimensions regarding rights and reasons to grant them. A potentially interesting approach for such future work would be to relate our findings to the more general literature on technology acceptance (e.g., to understand how experience with robots factors into attitudes of people (Turja and Oksanen, 2019)) or to compare the current reasons to grant robots rights and the mental capacities (Malle, 2019) revealing potential missing coverage in the reasons. Finally, future research should explore the effect of a robot's physical appearance on granting robots rights beyond the mechanical-humanoid dimension applied in our study.

5.2 Conclusion

Our study presents a survey design to empirically investigate the public opinion about robot rights. There appears to be an overall consensus about the interactive potential of robots. We found that people are more willing to grant basic robot rights such as access to energy and the right to update to robots than sociopolitical rights such as voting rights and the right to own property. We did not find any strong relation between demographic factors such as age or other factors such as experience with robots or of geographical region with the willingness to assign rights to robots. We did find, however, that beliefs about the (future) capacities of robots influence this willingness. Our results suggest that, in order to reach a broad consensus about assigning rights to robots, we will first need to reach an agreement in the public domain about whether robots will ever develop cognitive and affective capacities.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by VU Amsterdam. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contributions to the work, to the design of the survey, construction of the materials and instruments, and approved it for publication. MD and KH contributed to data-collection. MD contributed to data preparation and data-analyses. FH contributed to conceptual questions about reasons and rights. KH advised on the data-analyses and contributed most to the identification of the reason and rights items included in the survey. All authors jointly discussed and contributed to the final formulation of the items in the survey.

[9] 2019 Revision of World Population Prospects, United Nations, https://population.un.org/, accessed on September 3, 2021.

REFERENCES

Bartneck, C., Suzuki, T., Kanda, T., and Nomura, T. (2007). The Influence of People's Culture and Prior Experiences with Aibo on Their Attitude Towards Robots. *Ai Soc.* 21, 217–230. doi:10.1007/s00146-006-0052-7

Basl, J. (2014). Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. *Philos. Technol.* 27, 79–96. doi:10.1007/s13347-013-0122-y

Billing, E., Rosén, J., and Lindblom, J. (2019). "Expectations of Robot Technology in Welfare," in The Second Workshop on Social Robots in Therapy and Care in Conjunction with the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2019), Daegu, Korea, March 11–14 2019.

Block, N. (1995). On a Confusion About a Function of Consciousness. *Behav. Brain Sci.* 18, 227–247. doi:10.1017/S0140525X00038188

Borenstein, J., and Arkin, R. (2016). Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Sci. Eng. Ethics* 22, 31–46. doi:10.1007/s11948-015-9636-2

Bratchell, N. (1989). Cluster Analysis. *Chemometrics Intell. Lab. Syst.* 6, 105–125. doi:10.1016/0169-7439(87)80054-0

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: The Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25, 273–291. doi:10.1007/s10506-017-9214-9

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *Calif. L. Rev.* 103, 513–563. doi:10.2307/24758483

Chopra, S., and White, L. (2004). "Artificial Agents-Personhood in Law and Philosophy," in Proceedings of the 16th European Conference on Artificial Intelligence, Valencia Spain, August 22–27, 2004 (IOS Press), 635–639.

Ciepley, D. (2013). Beyond Public and Private: Toward a Political Theory of the Corporation. *Am. Polit. Sci. Rev.* 107, 139–158. doi:10.1017/S0003055412000536

Coeckelbergh, M. (2021). How to Use Virtue Ethics for Thinking About the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance. *Int. J. Soc. Robotics* 13, 31–40. doi:10.1007/s12369-020-00707-z

Coeckelbergh, M. (2010). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12, 209–221. doi:10.1007/s10676-010-9235-5

Colagrosso, M. D., and Mozer, M. C. (2005). "Theories of Access Consciousness," in *Advances in Neural Information Processing Systems 17*. Editors L. K. Saul, Y. Weiss, and L. Bottou (Cambridge, MA, USA: MIT Press), 289–296.

Danaher, J. (2016). Robots, Law and the Retribution Gap. *Ethics Inf. Technol.* 18, 299–309. doi:10.1007/s10676-016-9403-3

Darling, K. (2016). "Robot Law," in *Extending Legal protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects (April 23, 2012)*. Editors A. Ryan Calo, M. Froomkin, and I. Kerr (Glos, UK: Edward Elgar Publishing). doi:10.2139/ssrn.2044797

De Graaf, M. M., and Allouch, S. B. (2016). "Anticipating Our Future Robot Society: The Evaluation of Future Robot Applications from a User's Perspective," in International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, USA, 26–31 August, 2016 (IEEE), 755–762. doi:10.1109/roman.2016.7745204

de Graaf, M. M. A., and Malle, B. F. (2019). "People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences," in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, South Korea, March 11–14, 2019, 239–248. doi:10.1109/hri.2019.8673308

De Graaf, M. M., Hindriks, F. A., and Hindriks, K. V. (2021). "Who Wants to Grant Robots Rights," in International Conference on Human-Robot Interaction (HRI), Cambridge, UK (virtual), March 09–11, 38–46.

Fink, J. (2012). "Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction," in International Conference on Social Robotics, Chengdu, China, October 29–31, 2012 (Springer), 199–208. doi:10.1007/978-3-642-34103-8_20

Flandorfer, P. (2012). Population Ageing and Socially Assistive Robots for Elderly Persons: The Importance of Sociodemographic Factors for User Acceptance. *Int. J. Popul. Res.* 2012, 829835. doi:10.1155/2012/829835

Freitas, R. A. (1985). Can the Wheels of justice Turn for Our Friends in the Mechanical Kingdom? Don't Laugh. *Student lawyer* 13, 54–56.

Gerdes, A. (2016). The Issue of Moral Consideration in Robot Ethics. *SIGCAS Comput. Soc.* 45, 274–279. doi:10.1145/2874239.2874278

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of Mind Perception. *Science* 315, 619. doi:10.1126/science.1134475

Gray, K., Young, L., and Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychol. Inq.* 23, 101–124. doi:10.1080/1047840X.2012.651387

Gunkel, D. J. (2014). A Vindication of the Rights of Machines. *Philos. Technol.* 27, 113–132. doi:10.1007/s13347-013-0121-z

Gunkel, D. J. (2018). *Robot Rights*. Cambridge, MA, USA: MIT Press.

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA, USA: MIT Press.

Haslam, N., Bastian, B., Laham, S., and Loughnan, S. (2012). "Humanness, Dehumanization, and Moral Psychology," in *Herzliya Series on Personality and Social Psychology. The Social Psychology of Morality: Exploring the Causes of Good and Evil*. Editors M. Mikulincer and P. R. Shaver (Washington, DC, USA: American Psychological Association), 203–218. doi:10.1037/13091-011

Haslam, N. (2006). Dehumanization: An Integrative Review. *Pers Soc. Psychol. Rev.* 10, 252–264. doi:10.1207/s15327957pspr1003_4

Jaynes, T. L. (2020). Legal Personhood for Artificial Intelligence: Citizenship as the Exception to the Rule. *AI Soc.* 35, 343–354. doi:10.1007/s00146-019-00897-9

Kuehn, J., and Haddadin, S. (2017). An Artificial Robot Nervous System to Teach Robots How to Feel Pain and Reflexively React to Potentially Damaging Contacts. *IEEE Robot. Autom. Lett.* 2, 72–79. doi:10.1109/LRA.2016.2536360

Laukyte, M. (2019). "Ai as a Legal Person," in Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, Montreal, Canada, June 17–21, 2019 (New York, NY: Association for Computing Machinery), 209–213. doi:10.1145/3322640.3326701

Levy, D. (2009). The Ethical Treatment of Artificially Conscious Robots. *Int. J. Soc. Robotics* 1, 209–216. doi:10.1007/s12369-009-0022-6

MacDorman, K. F., Vasudevan, S. K., and Ho, C.-C. (2008). Does Japan Really Have Robot Mania? Comparing Attitudes by Implicit and Explicit Measures. *AI Soc.* 23, 485–510. doi:10.1007/s00146-008-0181-2

Malle, B. (2019). "How Many Dimensions of Mind Perception Really Are There," in Proceedings of the 41st Annual Meeting of the Cognitive Science Society, Montreal, Canada, July 24–27, 2019. Editors A. K. Goel, C. M. Seifert, and C. Freksa (Cognitive Science Society), 2268–2274.

Malle, B., and Thapa, S. (2017). *Unpublished Robot Pictures*. Providence, RI, USA: Brown University.

Matthias, A. (2004). The Responsibility gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics Inf. Technol.* 6, 175–183. doi:10.1007/s10676-004-3422-1

Miller, L. F. (2015). Granting Automata Human Rights: Challenge to a Basis of Full-Rights Privilege. *Hum. Rights Rev.* 16, 369–391. doi:10.1007/s12142-015-0387-x

Murtagh, F., and Legendre, P. (2011). Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *Stat* 1050, 11. doi:10.1007/s00357-014-9161-z

O'connor, B. P. (2000). Spss and Sas Programs for Determining the Number of Components Using Parallel Analysis and Velicer's Map Test. *Behav. Res. Methods Instr. Comput.* 32, 396–402. doi:10.3758/BF03200807

Phillips, E., Ullman, D., de Graaf, M. M. A., and Malle, B. F. (2017). What Does a Robot Look like?: A Multi-Site Examination of User Expectations about Robot Appearance. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 61, 1215–1219. doi:10.1177/1541931213601786

Regan, T. (2004). *The Case for Animal Rights*. Berkeley, CA, USA: Univ of California Press.

Schwitzgebel, E., and Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Stud. Philos.* 39, 98–119. doi:10.1111/misp.12032

Singer, P. (1974). All Animals Are Equal. *Philosophic Exchange* 5, 6.

Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016). The Relevance Effect and Conditionals. *Cognition* 150, 26–36. doi:10.1016/j.cognition.2015.12.017

Solum, L. B. (1992). Legal Personhood for Artificial Intelligences. *North Carolina L. Rev.* 70, 1231–1288.

Sullins, J. P. (2010). Robowarfare: Can Robots Be More Ethical Than Humans on the Battlefield? *Ethics Inf. Technol.* 12, 263–275. doi:10.1007/s10676-010-9241-7

Tavani, H. (2018). Can Social Robots Qualify for Moral Consideration? Reframing the Question About Robot Rights. *Information* 9, 73. doi:10.3390/info9040073

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Q.* 30, 415–433. doi:10.1177/107769905303000401

Tessier, C. (2017). *Robots Autonomy: Some Technical Issues*. Cham: Springer International Publishing, 179–194. doi:10.1007/978-3-319-59719-5_8

Tomarken, A. J., and Serlin, R. C. (1986). Comparison of Anova Alternatives Under Variance Heterogeneity and Specific Noncentrality Structures. *Psychol. Bull.* 99, 90–99. doi:10.1037/0033-2909.99.1.90

Torrance, S. (2012). Super-intelligence and (Super-)consciousness. *Int. J. Mach. Conscious.* 04, 483–501. doi:10.1142/S1793843012400288

Turja, T., and Oksanen, A. (2019). Robot Acceptance at Work: A Multilevel Analysis Based on 27 Eu Countries. *Int. J. Soc. Robotics* 11, 679–689. doi:10.1007/s12369-019-00526-x

van den Hoven van Genderen, R. (2018). *Do We Need New Legal Personhood in the Age of Robots and AI*. Singapore: Springer Singapore, 15–55. doi:10.1007/978-981-13-2874-9_2

Ward, J. H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* 58, 236–244. doi:10.1080/01621459.1963.10500845

Weisman, K., Dweck, C. S., and Markman, E. M. (2017). Rethinking People's Conceptions of Mental Life. *Proc. Natl. Acad. Sci. USA* 114, 11374–11379. doi:10.1073/pnas.1704347114

Wellman, C. (2018). *The Proliferation of Rights: Moral Progress or Empty Rhetoric*. New York, NY, USA: Routledge.

Wiener, N. (1954). *The Human Use of Human Beings: Cybernetics and Society*, 320. Boston, MA, USA: Houghton Mifflin.

Wilkinson, C., Bultitude, K., and Dawson, E. (2011). "oh Yes, Robots! People like Robots; the Robot People Should Do Something": Perspectives and Prospects in Public Engagement with Robotics. *Sci. Commun.* 33, 367–397. doi:10.1177/1075547010389818

Wootson, C. (2017). Saudi Arabia, Which Denies Women Equal Rights, Makes a Robot a Citizen. *The Wash. Post.*

Złotowski, J., Proudfoot, D., Yogeeswaran, K., and Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *Int. J. Soc. robotics* 7, 347–360. doi:10.1007/s12369-014-0267-6

Zygmont, C., and Smith, M. R. (2014). Robust Factor Analysis in the Presence of Normality Violations, Missing Data, and Outliers: Empirical Questions and Possible Solutions. *Quantitative Methods Psychol.* 10, 40–55. doi:10.20982/tqmp.10.1.p040

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

SYSTEMATIC REVIEW
published: 11 January 2022
doi: 10.3389/frobt.2021.788179

Gradient Legal Personhood for AI Systems—Painting Continental Legal Shapes Made to Fit Analytical Molds

Diana Mădălina Mocanu *

Centre for Philosophy of Law (CPDR), Institute for Interdisciplinary Research in Legal Sciences (JUR-I), Université Catholique de Louvain, Louvain-la-Neuve, Belgium

What I propose in the present article are some theoretical adjustments for a more coherent answer to the legal "status question" of artificial intelligence (AI) systems. I arrive at those by using the new "bundle theory" of legal personhood, together with its accompanying conceptual and methodological apparatus as a lens through which to look at a recent such answer inspired from German civil law and named *Teilrechtsfähigkeit* or partial legal capacity. I argue that partial legal capacity is a possible solution to the status question only if we understand legal personhood according to this new theory. Conversely, I argue that if indeed *Teilrechtsfähigkeit* lends itself to being applied to AI systems, then such flexibility further confirms the bundle theory paradigm shift. I then go on to further analyze and exploit the particularities of *Teilrechtsfähigkeit* to inform a reflection on the appropriate conceptual shape of legal personhood and suggest a slightly different answer from the bundle theory framework in what I term a "gradient theory" of legal personhood.

Keywords: AI systems, bundle theory of legal personhood, partial legal capacity, *teilrechtsfähigkeit*, gradient theory of legal personhood

OPEN ACCESS

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Didem Gurdur Broo,
Royal Institute of Technology, Sweden
Anne Gerdes,
University of Southern Denmark,
Denmark

Correspondence:
Diana Mădălina Mocanu
diana.mocanu@uclouvain.be

Specialty section:
This article was submitted to
Ethics in Robotics and Artificial
Intelligence,
a section of the journal
Frontiers in Robotics and AI

Received: *01 October 2021*
Accepted: *18 November 2021*
Published: *11 January 2022*

Citation:
Mocanu DM (2022) Gradient Legal
Personhood for AI Systems—Painting
Continental Legal Shapes Made to Fit
Analytical Molds.
Front. Robot. AI 8:788179.
doi: 10.3389/frobt.2021.788179

PROLEGOMENA

"One cannot be too careful with words; they change their minds just as people do," Nobel Prize–winning writer Jose Saramago once warned. "Words are events; they do things, change things," Ursula le Guin, another distinguished writer and one of his devoted admirers, added. Nowhere else is this truer perhaps than in law's empire[1], where words such as "person" and "thing" are code for a "legal regime." That is, they have the power to trigger a host of consequences once applied. In order to apply them, jurists qualify and categorize reality, thus establishing links between what is and what ought to be. The trouble with this attempt at fighting entropy by conceptually ordering reality is that the latter sometimes simply refuses to play by the rules that we set. This means that new entities in the world do not always fit our existing legal molds, and so we are faced with a conundrum: do we create new molds, or do we tweak and twitch (our understanding of) our entities to fit the old ones?

[1] To echo Ronald Dworkin in *Law's Empire*, Belknap Press, 1986, whose fictitious judge Hercules would perhaps be ideally situated to undertake the work of legal interpretation and hermeneutics required to properly situate new technologies in the legal domain.

These very old such molds are currently being stretched to fit one very new type of entity, namely, AI systems[2]. Within the technical, as within the legal realm, word choice is crucial. Previously dubbed "autonomous artificial agents" or AAAs for short (Chopra and White, 2011) or "mathematically formalized information flows" (Teubner, 2018) and based on "artificial intelligence" (AI), to use a "suitcase word"[3], AI systems exploit myriad increasingly complex and oftentimes opaque methods found at the intersection of computer science, statistics, and other fields to enable solutions that are adaptive in order to perform specific tasks with a degree of autonomy.

What all AI systems have in common is a set of features that make them straddle the border between "thing" and "person." These are somewhat disputed, but most sources cite autonomy, usually related to a measure of agency, then adaptivity and self-learning. Embodiment is sometimes added to the list with claims that "genuine intelligence can emerge only in embodied, situated, cognitive agents" (Menary, 2007; Clark, 2017). This is lucky for robots but excludes a whole range of software-based entities. Adaptivity and self-learning abilities are sometimes reunited under "intelligence" (Kiršienė et al., 2011), another suitcase word with little to offer us in the way of a definition and perhaps even less so ever since it was coupled with "artificial" at Dartmouth in the 1950s. As it turns out, perhaps the most lasting contribution of this first attempt was not scientific, but semantic[4].

The phrase "artificial intelligence" brings up questions about whether intelligent behavior implies or requires the existence of a (human) mind or to what extent consciousness, if real, is replicable as computation. Legally, properties such as consciousness have traditionally[5] served as "qualifying criteria" (Gunkel and Wales, 2021) for natural personhood as opposed to the "artificial personhood" (Dyschkant, 2015) of entities such as corporations, which is said to be born out of the practical consideration of furthering some human interest more effectively. The term "intelligent machine" has recently been proposed as a metaphor for understanding both corporations and AI systems (Laukyte, 2021). Although somewhat unhelpful in cutting the Gordian knot due to the fact that we lack a definition for both intelligence or consciousness and that we have "epistemological limitations" (Gunkel and Wales, 2021) as to their detection in others, words and phrases such as these have been used by the law nonetheless, their vagueness making them the most debated points of contention in fringe cases on issues such as corporate rights, abortion, or euthanasia.

One of the few certainties we have is that AI systems are already widely used and will most likely infiltrate more and more aspects of everyday life, causing not just the way people think to be affected, but also the way they act and the manner in which they behave in their private and professional lives. This makes them legally significant because case-law that "anticipates the legal principles that may come to govern displacement of human activity by intelligent artifacts" (Wein, 1992) is bound to follow.

Because consensus is in general lacking as to whether legal innovation is in order, however, legal scholarship identified three criteria to determine when a new law is needed (Hondius, 1980). First, the legal problem has to fall either outside the scope of any existing branch of law or simultaneously under several branches, none of which resolves all aspects of the problem. Second, it has to affect broad sections of society and be likely to persist for a long time. Third, the new law has to result in basic principles sanctioned by the constitutional and legal system of the country concerned. Although this last one might prove problematic for legal innovations, because they mean changes to the very constitutional and legal systems by which they are supposed to be sanctioned, all three conditions are arguably met by AI systems' legal "status question" (Papakonstantinou and de Hert, 2020).

THE STATUS QUESTION

The status question (Schirmer, 2020) asks what exactly AI systems are, legally speaking. It makes us ponder whether we are just looking at sophisticated objects or things, whether we would rather treat them as legal persons, somewhat similar to humans or corporations, or indeed whether we should create a new legal category suited to their specificities. This way of phrasing the question mixes what exists in a material sense with what ought, from a moral point of view, to be and what we conventionally decide is or will legally be the case. Aside from the fact that we must exercise great care in juggling registers, because it is logically unsound and morally hazardous to derive an "ought" from an "is" (Norton and Norton, 2007) and slip from factual to axiological statements, this also reenacts to an extent the positivism versus natural law debate in the philosophy of law. Without becoming embroiled in the larger moral debate about AI systems as moral agents or patients, this

[2] I use the term "AI systems" throughout this paper, borrowing it from the European Commission's Proposal for a Regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), fully aware of the criticism directed at the too broad definition thereof. This choice is motivated by the fact that the author of the bundle theory of legal personhood expressly avoids reference to AI systems as "agents" or "actors" or even "actants" to underline the fact that legal personhood can be purely passive as with infants. More details on passive legal personhood can be found in **section 5**.

[3] Marvin Minsky, American mathematician, computer scientist, and AI pioneer, in an interview on the subject of his book *The Emotion Machine*, Simon and Schuster, 2006, conducted by John Brockman for Edge.org and titled *Consciousness is a big suitcase: A Talk With Marvin Minsky*, available at: https://www.edge.org/conversation/marvin_minsky-consciousness-is-a-big-suitcase. Interestingly, in his previous book, *The Society of Mind*, Simon and Schuster Paperbacks, 1988, he used "agents" to refer to the units of his "society of mind," the metaphor employed to explain how intelligence and a mind are accretions of different combinations of rather unintelligent and mindless composing units.

[4] Roberts, J. (2016). Thinking Machines: The Search for Artificial Intelligence - Does History Explain Why Today's Smart Machines Can Seem So Dumb? Distillations magazine, Summer issue. Available at: https://web.archive.org/web/20180819152455/https://www.sciencehistory.org/distillations/magazine/thinking-machines-the-search-for-artificial-intelligence.

[5] Recently, however, support against the idea that the criteria of legal subjectivity are consciousness, sentience or reason has been amassing. See for example Gunther Teubner, *op.cit.,* who proposes communication as the salient criterion while citing numerous other criteria proposed; for an even more recent contribution see Sylwia Wojtczak, *op.cit.*, proposing participation or presence in social life as the salient criterion.

contribution is strictly limited to law as artifactual, whether in order to refine expressions of some perfect idea of law (as natural law proponents hold) or simply to express renewed conventions (as positivists do).

The stark opposition between the two did not always exist as such though, and there might be a way around it too. Legal history shows us that legal fluidity coexists until it is gradually replaced by positive law, which is not immutable, but has to be adjusted to changes in society[6] to not become obsolete or, worse yet, unjust, which makes the cycle repeat itself. An early but evocative example of this process would be the usage of *jus* by later Romans, whereas Cicero said *lex.* This is significant because the ambiguity of *jus* "lending itself to identification of what ought to be and what is, gave a scientific foundation for the belief of the jurisconsults that when and where they were not bound by positive law they had but to expound the reason and justice of the thing in order to lay down the law" (Pound, 1922). That is, natural law is an approach best suited for times of change, when jurists need to use their judgement to make analogies and, when that does not suffice, to create law to apply to new social realities. It is unsurprising in this light that a case for a natural law conception of AI legal personhood was made and assessed in the context of contemporary legislative proposals, concluding, however, that the time for creating such a concept is not ripe yet (Jowitt, 2021).

Creative periods of fluidity in legal history generally follow stable ones. As things stand today, it is difficult to find a more constant and undisputed legal assumption than the one underlying the conceptual framework of juridical humanism, widely accepted in Western legal systems and which rests on the dualistic division of legal reality into persons and things. I argue that this model of the world is an oversimplified one and that there is general skepticism to more inclusive change. This is beginning to change, however, since juridical humanism has been criticized as incoherent (Pottage and Mundy, 2004; Pietrzykowski, 2017), requiring a reconceptualization and reorganization of the relationships between these and new categories.

This incoherence stems from historical exclusions from the category of legal person of women and slaves and the inclusion of fringe cases, such as newborn children, differently abled adults, or animals (which is incongruous with the traditional definition of legal personhood) as well as (putative) attributions of personhood to rivers, idols, ships. For example, numerous European legal systems now explicitly exclude animals from the category of things, but there is no language as to a new category they may be part of although several suggestions exist, including "nonpersonal subjects of law" (Pietrzykowski, 2017) or "nonhuman (natural) persons" (Regad and Riot, 2018; Regad and Riot, 2020). Instead, positive law likens their treatment to that of goods, prompting legitimate complaints from animal rights activists about the purely formal "change" in status that practically amounts to as little protection and participation in legal life as before.

Change in the legal status of animals, let alone AI systems, is "not simply unacceptable, but rather unthinkable for many jurists" (Kurki, 2019)[7]. This is because the divide between persons and nonpersons is a part of the "deep structure of law" (Tuori, 2002), and questioning that binomial relationship is no easy feat. It has ample practical value though, and I argue it can be best accomplished through the logical and orderly analysis of law, which is the prime mission of the philosophy of law, or at the very least it is in its analytical bent. It bears noting at this point that the continental-analytical juxtaposition in the title can be misleading given that "continental" is used to refer to continental legal systems, and more specifically civil law, and not continental philosophy, whereas "analytical" does refer to the homonymous philosophical tradition. More specifically, continental legal shapes refer to the legal concepts of person and thing such as they exist in civil law traditions on the European continent. Analytical molds refer to the coherence of such legal concepts according to analytical methods.

Historically, philosophy "has been used to break down the authority of outworn tradition, to bend authoritatively imposed rules that admitted of no change to new uses which changed profoundly their practical effect, to bring new elements into the law from without and make new bodies of law from these new materials, to organize and systematize existing legal materials and to fortify established rules and institutions when periods of growth were succeeded by periods of stability and of merely formal reconstruction" (Pound, 1922). A method for such innovation has also already been proposed in relation to AI[8], namely, conceptual engineering (Chalmers, 2020; McPherson and Plunkett, 2020). It holds that clarifying the content of core concepts should be the first step of any debate to avoid arguing about different things, but also to recover conceptual possibilities by figuring out what our concepts actually stand for and, more importantly, what they ought to stand for. We may, however, need to first engineer the concept of conceptual engineering itself, which is procedurally far from clear. This is in order to avoid a recursive engineering loop—a somewhat

[6] Pound offers a very early example of such coexistence, which comes to us in the form of an exhortation addressed by Demosthenes to an Athenian jury, saying that "men ought to obey the law for four reasons: because laws were prescribed by God, because they were a tradition taught by wise men who knew the good old customs, because they were deductions from an eternal and immutable moral code and because they were agreements of men with each other binding them because of a moral duty to keep their promises." Modern legal eyes might just dismiss these four reasons as contradictory, but that would be ignoring the fact that they served a very practical purpose of establishing social control in a primitive society by whatever means necessary. The classical Greeks were just then trying to cement some basis of authority for law, which we today largely take for granted although we still question whether right is right by nature or by convention. That would also be ignoring another factor in the establishment of social control through law, namely, that it takes time.

[7] David Gunkel has recently mapped where authors having written on the subject of robot rights fall on that debate, with unthinkable associated to names such as Noel Sharkey, Luciano Floridi, Alan Winfield, Sherry Turkle, Abeba Birhane, or Jelle van Dijk.

[8] Köhler, S., and Himmelreich, J. (2021). Responsible AI through Conceptual Engineering, Talk Given for TUDelft's. AiTech Agora.

fitting irony given the context of application to AI systems—of both concept and method.

It is at any rate becoming increasingly urgent for law to take a stand in answering the status question for the case of AI systems. To do otherwise is to allow the possibility of "potentially impeding further development and the practical usefulness of the whole technology" (Pietrzykowski, 2017). Giving voice to the law on these matters should, however, aim to maintain or indeed establish the coherence of its discourse throughout. Therefore, in the interest of walking a moderate path, we would be ill-advised to legally tip the balance of power characteristic of the politics of nature (Latour, 2004) so irrevocably in our favor in relation to AI systems as to assume absolute responsibility for them (Kruger, 2021) without due consideration to the general question of whether AI systems should be regarded as being in our service or rather if the circle of legal subjects should be enlarged instead so as to include nonhuman entities.

To this question the EU seems to offer in answer its human-centric approach to AI regulation[9]. That law is an anthropocentric construction is fairly undisputed. It should perhaps come as no surprise that we have been tipping the scales of justice in our favor all along, in light of this premise. Human beings have made law preoccupied first and foremost with themselves and their wishes as to ordering lived reality. In an overt admission of speciesism, it is claimed (Bryson et al., 2017) that the main purposes of any human legal system revolve around giving preference to human material interests as well as human legal and moral rights and obligations over the similar claims of any nonhuman entities.

CURRENT ANSWERS TO THE STATUS QUESTION

Whether AI systems could be accorded (some form of) legal personhood, thus entering law's ontology as legal persons is "a matter of decision rather than discovery" (Chopra and White, 2011). The same is true for qualifying AI systems as things however, and we gather that much, at least at a declarative level, from the EU Parliament's apparent change of tune from the creation of "electronic persons" in 2017[10] to its 2020 Resolution[11]. There is currently "no need" to give a legal personality to emerging digital technologies we are told in the latter. The initial ambition for a paradigm shift manifested in 2017 (Sousa Antunes, 2020) is, thus, wholly missing from more recent documents and that despite the fact that "moving past an anthropocentric and monocausal model of civil liability" was seen as a potential "unifying event" in a report[12], which the commission had shortly before it tasked its Expert Group on Liability and New Technologies.

Subsequently, such a solution was not, however, seen as practically useful, mainly because "civil liability is a property liability, requiring its bearer to have assets" in order to give "a real dimension" to it, which would, in turn, require "the resolution of several legislative problems related to their legal capacity and how they act when performing legal transactions." Despite acknowledging the fact that giving AI systems legal personality would not require including all the rights that natural or even legal persons have and that, theoretically, their legal personality could consist solely of obligations, such a step was still considered too much of a leap. It might well be, as opposed to just tinkering with traditional (liability) solutions while engaging in wishful thinking as to the harm caused by these technologies being reducible to risks that can ultimately be attributed to existing, albeit unidentifiable, natural or legal persons.

In the likely event of such attribution of harm to (legal) persons being hampered by too complex production chains or breaks in the chain of causality, affected parties do not have sufficient and effective guarantees of redress. Instead, we are simply told that "new laws directed at individuals are a better response than creating a new category of legal person" (Abbott and Sarch, 2019). Indeed, the tendency to rely on existing interpretations of the law instead of innovating is apparent in the European Commission's Proposal for an AI Act as well (Veale and Zuiderveen, 2021).

The fact remains that humans or, shall we say, the natural persons who are taking the status decision, are inevitably bound to bias the answer toward safeguarding some human interests, especially given that AI systems currently do not (fully) display features traditionally considered as salient for the attribution of interests of their own. Which human interests get to be safeguarded is a balancing act that, for example, in the text of the 2020 Resolution arguably gave way to AI systems' operators' and emerging AI industries' interests by softening risk-based liability. It introduced a two-tiered system of liability (with strict liability for high-risk systems and subjective liability with a presumption of fault for ordinary-risk systems) where only strict liability used to apply before. Indeed the creation and use of strict liability, or liability without fault, is linked with technological progress. What is more, the same text repeatedly warns against the overall increased risks involved in the operation of AI systems, which involve loss in control on the part of human operators. It seems counterintuitive in this light that "an increase in the risk factors indicated weakens the liability of the actor"

[9] The term is part and parcel of the recent European Commission Proposal for an AI Act mentioned previously, published on April 21, 2021 and available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788.

[10] European Parliament (2017) Resolution of 16 February with recommendations to the Commission on Civil Law Rules on Robotics [2015/2103(INL), sec. 59 f] states that "in the long run, it will be necessary to create a specific legal status (. . .), so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently."

[11] European Parliament, Report with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)), available at: https://www.europarl.europa.eu/doceo/document/A-9-2020-0178_EN.html.

[12] Expert Group on Liability and New Technologies, New Technologies Formation, *Liability for Artificial Intelligence and other emerging digital technologies*, p. 19, available at: https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608.

(Sousa Antunes, 2020). Moreover, producers' liability may not be amenable to similar differentiation into a two-tiered system, and while the fact that the two (i.e. operators' and producers' liability) need to articulate well for a liability regime to function well is undisputed, the question of how this articulation could function at all given this novel asymmetry remains unanswered as yet.

These mark only the beginning of a long string of pieces of legislation likely to tackle AI systems, however, and it would be premature to say that there has been a departure from the 2017 paradigm shift to the point of no return, especially given the phrasing "in the long run" utilized then to refer to the time frame of granting some form of legal personhood to "at least the most sophisticated robots" in order for them to "make good any damage they may cause"[13]. It has been argued (Papakonstantinou and de Hert, 2020) that liability is enhanced, not reduced, through granting legal personality to AI, the first and most important advantage of that being flexibility for every branch of law to assess the legal issues posed by AI systems within its own boundaries and under its own rules and principles, leading to tailor-made solutions as opposed to a "supervisory authority" with an opaque legal mandate to "monitor" any and all AI systems. Another advantage would be the proximity of one-to-one legal relationships with AI systems instead of the multitude of stakeholders involved in creating, operating, or putting them on the market, which, given modern production chains, are likely scattered all over the globe and prohibitively complex (Crawford, 2021).

The current piecemeal approach to regulating AI in the EU by identifying the sectors most likely to be affected by AI, highlighting potential problems and making concrete punctual suggestions for legislative intervention in order to address them "is in effect an amendment through *ad hoc* patches" (Papakonstantinou and de Hert, 2020) of the legal framework currently in effect using existing legal tools. It might amount to a change in legal status nonetheless, given enough tinkering, but a formal recognition of that would still need to come either *via* case law decided by the Court of Justice of the European Union or positive law.

THE CASE FOR LEGAL CREATIVITY

The legal nature of AI systems is a preoccupation arising within "legal reality" (Hermitte, 1999), an environment that purports to organize and, thus, help make sense of lived reality, in which technology evolves in ways that challenge our legal models. We are, thus, faced with having to breach a gap that, in the long run, will presumably only deepen. Pragmatically, their functionality and social role as well as our relationships with AI systems will probably be the decisive arguments to sway the answer to the status question. The economic context was even said to lead to changes in the status of AI systems before that of nonhuman animals (Michalczak, 2017).

Arguing that AI systems may have potential legal subjectivity based on an analogy to animals, however, or even juristic persons for that matter, superficially suggests "the existence of a single hierarchy or sequence of entities, organized according to their degree of similarity to human beings" (Wojtczak, 2021). The place of an entity in this hierarchy would determine the scope of subjectivity attributed to it, a subjectivity that would be "derivative" in nature and not different from that of animals and companies.

Subjecthood could instead become a sort of master mold. Diversifying status thus, we would create the all-encompassing meta-category of subjects, including persons and other "nonpersonal subjects" (including human nonpersonal subjects and extra-human nonpersonal subjects). Nonhuman animals were already used as an example of beings whose legal status could be changed from things to "nonpersonal subjects"—not quite legal persons, but not things either (Pietrzykowski, 2018). Such subjects would, according to this opinion, differ from traditional persons in that they would be the holders of limited rights, or—in the case of animals—the single subjective right to be taken into account and have their interests duly considered and balanced whenever legal decisions affect them.

To further complicate matters, in many European languages, the term or phrase "legal subject" or "subject of law/right(s)" (*Rechtssubjekt*; *sujet de droit*) already is an umbrella term, referring to both natural and artificial persons, i.e., individual human beings and corporations or other such associations, respectively. This usage of *Rechtssubjekt* was introduced by Savigny. Civil law jurisdictions use the phrase "legal subject" or "subject of law" when addressing legal persons, whereas such phrases may seem odd to common lawyers, who use legal person to refer to artificial persons (Kurki and Pietrzykowski, 2017).

This begs the question of how exactly a Venn diagram might show the relationships between these concepts. We could for instance, looking at the diagram in **Figure 1**, imagine the circle of "things" intersecting that of "non-personal legal subjects" and therefore the larger "legal subjects" one. It also allows us to question in which of the categories illustrated below AI systems might end up included, or indeed whether an entirely new category might be defined specially to include them.

What thus becomes paramount is delimiting the extension of the concept of legal personhood, the attribution of which has in legal theory been thought of as either requiring certain preexisting conditions, or not requiring any at all. If it does not, then it is in this case merely a fiction[14], an instrument of law or a label that we apply to trigger certain consequences, which can be either

[13] European Parliament, Resolution of February 16, 2017, with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

[14] Variations of this view are called legalist, with the so-called "anything-goes approach to legal personhood" attributable to Natalie Stoljar by Ngaire Naffine in *Who are Law's Persons? From Cheshire Cats to Responsible Subjects*, 66 Modern Law Review, 2003, pp. 346–351.

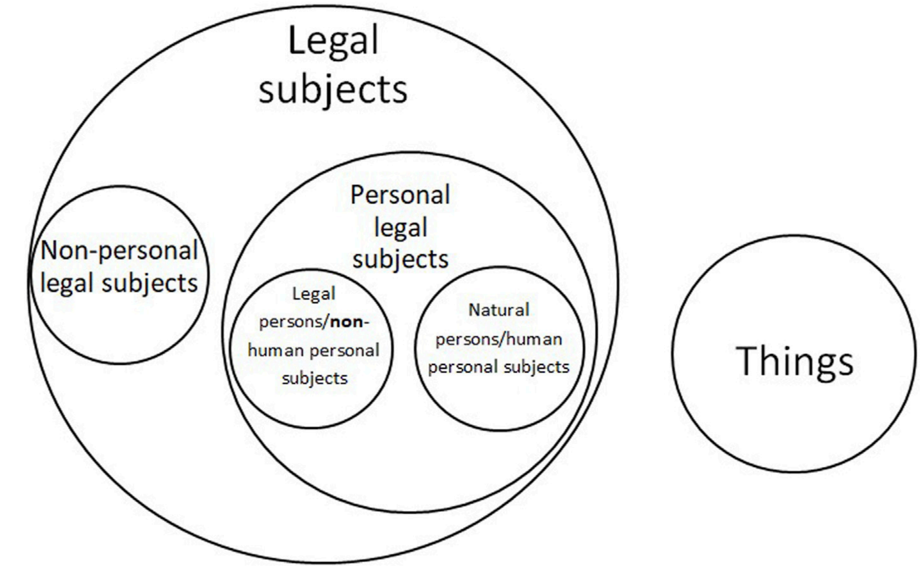
Legal subjects
Non-personal legal subjects
Personal legal subjects
Legal persons/non-human personal subjects
Natural persons/human personal subjects
Things

FIGURE 1 | Diversifying legal status.

predetermined in bulk or established on a case-by-case basis. If it does, however, require certain preexisting conditions, then the debate moves to determining what these may be. Needless to say, there is no consensus.[15]

Either way, translating what is into what is in legal terms so that we can compare it against what we think ought to be and act on that basis to establish or restore some sense of balance is what jurists of all times have trained to do. It is also what they obsess about to such a degree that the is–ought distinction becomes blurred at times. To repeat a rather amusing metaphor for this confusion, "when a jurist gets bitten by a dog, they do not scream, but think about whether the conditions for liability for the animal's action are met" (Rizoiu, 2020). What if, instead, the jurist suffers harm as a consequence of the behavior of Boston Dynamics' dog-like robot, Spot? It was argued that facilitating liability mechanisms for holding AI systems directly liable for (at least some of) the effects of their decisions and actions is one of the most compelling arguments in favor of granting "sufficiently complex" AI systems "some form of personhood instead of regarding them as ordinary things (mere machines)" (Solum, 1992; Chopra and White, 2011). Whereas the possibility of applying "electronic personality" to them has been vastly criticized[16], the critiques going in this direction generally rely on a concept of legal personhood that is—it has been successfully claimed, as we shall see—in need of a reappraisal (Kurki, 2019).

LEGAL PERSONHOOD *QUA* BUNDLE

The bundle metaphor is used to depart from the "orthodoxy" (Kurki, 2019) of legal personhood as the capacity to hold rights and duties, explaining it instead as a cluster of interconnected "incidents." As a "cluster property" or a property "whose extension is determined based on a weighted list of criteria, none of which alone is necessary or sufficient"[17], legal personhood could have different configurations, mirroring different legal contexts. Legal personhood cannot be equated with the holding of rights because modern theories of rights, which are based on Hohfeld's conceptual clarifications on the notion of "right"[18], "either ascribe rights to entities that are not usually classified as legal persons, such as foetuses and nonhuman animals, or deny rights to entities that are ordinarily classified as legal persons, such as human children" (Kurki, 2019). There are glaring discrepancies between the list of holders of rights and obligations according to contemporary theories about the foundations of subjective rights and the list of persons according to the much older "orthodox view" on legal personhood, although they should be identical if the latter had an adequate definition.

Because "paradigmatic doctrinal judgements" and "extensional beliefs" about who or what constitutes a legal person would be nearly impossible to change and would, even

[15] We owe the legalist realist distinction to Ngaire Naffine too. Reason has ample support as the salient feature, but, religionists tell us it is the soul, and naturalists deem it to be sentience. All of these are variations of the so-called realist view.

[16] Open letter to the European Commission Artificial Intelligence and Robotics, signed by almost 300 professionals and experts in the relevant fields, available at: http://www.robotics-openletter.eu/.

[17] See John R. Searle, *Proper Names*, (1958) 67 *Mind* 166. What *is* necessary and sufficient is a disjunction of certain proper subsets of the set of cluster properties.

[18] Wesley Newcomb Hohfeld's writings are few because of his too short life, but nonetheless important contributions to legal theory: *The Relations Between Equity and Law*, 1913, Michigan Law Review, 537; *Some Fundamental Legal Conceptions as Applied in Legal Reasoning*, Yale Law Journal vol. 23, no. 1/1913, pp. 16-59; *Fundamental Legal Conceptions as Applied in Judicial Reasoning*, Yale Law Journal vol. 26, no. 8/1917, pp. 710-770.

if changed, offer little in the way of an explanation, what follows is that the definition of legal personhood should be adapted to accommodate modern theories of rights (Kurki, 2019). The "bundle theory" offers one such adaptation, using reflective equilibrium as a method[19].

For the specific case of AI systems, the bundle theory is used to analyze three contexts that influence the outcome of the debate over their legal status, namely "the ultimate value context" (whether artificial intelligence has intrinsic moral value, not derived from its usefulness to other entities), "the liability context" (whether artificial intelligence can be held tortuously or criminally liable for its actions), and last but not least "the commercial context" (whether artificial intelligence can function as a commercial actor).

The first of these three, ultimate value, is connected to passive legal personhood, which "functions through claim-rights" (Kurki, 2019), allowing correlative duties to be owed to AI systems. Even if an AI is of ultimate value, we are left without an explanation as to why this is relevant for status ascription *per se* so much as with the impression of the question being moved to the territory of ethics. If an AI is not of ultimate value, however, and this seems to be the case according to the bundle theory, then it can only hold claim rights as the administrator of a human-defined project. The only avenue left for them to be legal persons in that case runs through their "capacity to be subjected to legal duties and/or to administer legal platforms through the exercise of competences" (Kurki, 2019).

A "legal platform" refers to the legal positions held by a "legal person" and was introduced in the bundle theory to distinguish the two as well as to counterbalance the systematically ambiguous current doctrinal use of "legal person" to refer to both. That AI systems can hold claim rights as administrators of legal platforms with objectives set by human beings was concluded elsewhere too (Bayern, 2021), using similar examples to those utilized here, of an investment bank that hires trader bots to buy and sell stocks at a superhuman pace and a foundation run by an entity based on artificial intelligence.

In the commercial context, another distinction is superposed, between independent and dependent legal personhood based on the amount of supervision necessary in the exercise of competences by AI systems, which would place them on a tool–representative–legal person continuum.

Attaching legal platforms to entities that do not fulfill certain criteria would prove useless according to the bundle theory. The ability to hold claim rights makes for passive legal personhood and the ability to perform legal acts for active legal personhood. Passive legal personhood designates legal capacity and is likened to *Rechtsfähigkeit* or *capacité de jouissance* in German and French continental law, respectively, whereas active legal personhood corresponds to legal competence, *Geschäftsfähigkeit* or *capacité d'exercice*. These parallels to civil law traditions prove useful for analyzing a recently proposed solution to AI systems' legal status question, namely, *Teilrechtsfähigkeit* or partial legal capacity (Schirmer, 2020).

THE PARTIAL LEGAL CAPACITY VARIATION

This ontological category of legal subjects, halfway between person and object[20] was inspired from German civil law and dubbed "a half-way status" or "a status of partial legal subjectivity based on certain legal capabilities." Partial legal capacity would entail treating AI systems as legal subjects as far as this follows their alleged function of "sophisticated servants" (Schirmer, 2020).

Juridical humanism's all-or-nothing version of legal personhood is ill-suited for explaining such flexibility, which, in turn, seems to confirm the bundle theory. Born out of a critique of the two-tier system of legal capacity as inconsistent with the reality of how legal systems treat minors or used to treat women and slaves, partial legal capacity is a later materialization of the conclusion that legal capacity comes in plurals and there are, accordingly, many legal statuses. Defined in the 1930s as a status applicable to a human or an association of humans having legal capacity only according to specific legal rules but otherwise not bearing duties and having rights, it is, thus, an expansion of our understanding of legal capacity.

Although bent out of form and used for the practical disenfranchisement of the Jewish population[21], it survived *via* court judgments regarding the unborn or preliminary companies. In German law, the preliminary company (*Vorgesellschaft*) is considered a legal entity of its own kind (*Rechtsform sui generis*) subject only to the rules of the articles of association and the statute governing the company, insofar as those laws do not require registration. This also applies to certain company types such as the company constituted under civil law or the homeowner's association.

In the case of the first two, *i.e.,* the unborn and preliminary companies, the use case covered by partial legal capacity seems to be concerning entities "in the making." In this sense, it is a transitional state. The temporal and temporary dimension is more evident in some civil law jurisdictions than in others. For example, in Romanian civil law, the preliminary company enjoys "anticipatory legal capacity"[22] or limited legal capacity to perform the necessary legal acts in anticipation of its own formation. Article 60 of the Belgian *Code des sociétès* on the other hand sets an "imperfect liability," meaning that natural persons acting on behalf of the company (such as the founders) engage its personal liability in performing acts necessary for

[19] Known to us from John Rawls' writings.

[20] See Peter H. Kahn, Jr., et al., *The New Ontological Category Hypothesis in Human-Robot Interaction*, 2011 PROC. 6TH INT'L CONF. ON HUMAN-ROBOT INTERACTION 159; Ryan Calo, *Robotics and the Lessons of Cyber law*, 103 Calif. L. Rev. 513 (2015), https://digitalcommons.law.uw.edu/faculty-articles/23.

[21] According to Schirmer, this view of legal capacity as plural and governed only by specific legal rules that do not give rise to rights and obligations served as justification for the gradual subtraction of rights from the Jewish population leading up to the second World War.

[22] Article 205, (3), Romanian Civil Code states (*trad.n.*), "However, the legal persons dealt with in paragraph (1) (subject to registration) can, starting at the date of their constitutive act, acquire rights and take on obligation, but only insofar as they are necessary for the valid creation of said legal persons."

founding the company, such as renting an office. Once a legal person has been created, it takes on the contract itself, in a postconstitutive transfer of full liability. Here, the coherence of the institution of contracting is at stake. The use of the term "imperfect" denotes the same transitional state mentioned prior, which the law struggles to accommodate.

In the case of the latter examples, i.e., the company constituted under civil law or the homeowner's association, it is a question of the specific assortment of rights and duties attributable to a human or an association of humans having legal capacity only according to specific legal rules. This might refer perhaps to the law's presuppositions about the legal person in its various subdomains as it does about the diligence and reasonableness of the *bonus pater familias* wearing their administrator *persona* and which points, in turn, to an intrasystem asymmetry (Novelli et al., 2021) as to the meaning of personhood in different legal subfields. It might, on the other hand, refer to the limiting principle of specialization that circumscribes the legal capacity of juristic persons such as companies around its object as formulated in its statutes. The problem with this is that, in general, statute formulations are so vague and encompassing as to prevent any legal challenge based on an alleged *ultra vires* act performed in the company's name.

In terms of partial legal capacity, accretions of rights need to be justified according to the function of the entity in question, and the only binding expression of that function is, in the case of companies, their statute. In the case of AI systems, as we shall see, functionality is largely inscribed in the artifact but should also be formalized to avoid misuse or abuse. A way of inventorying their functions and necessary capacities to accomplish such functions so that abuse is kept in check would be *via* registries. Preliminary companies, however, must not be subject to registration if they are to possess partial/anticipatory/imperfect legal capacity and be considered legal entities of their own kind, which is to say that humans decide what legal persons they inventory.

At any rate, partial legal capacity does not work by limiting capacity, but by allocating or adding legal capacities as they are justified, as opposed to legal personhood, which asks us to justify their subtraction. This is how partial legal capacity is supposed to, solve the slippery slope of having to justify denying worker and constitutional rights to AI systems, which is one of the "negative side effects of full legal personhood" being attributed to these entities (Schirmer, 2020). Seen through the lens of the bundle theory and the above examples, partial legal capacity could actually amount to personhood, albeit as a smaller bundle.

These examples do not show legal persons with full legal capacity, but they do show legal subjects nonetheless, though with the range of their subjectivity limited by their specific functions. This characterization joins the bundle theory's assertion that there are several ways in which the law might treat entities in the world "more or less as persons" (Kurki, 2019). It might do so for a particular purpose and not others, it adds, pointing to the general variety of the law's purposes and the corollary flexibility required of legal personhood for it to better suit them. It leaves some doubts, however, as to the nature of the conceptual relationship between function, purpose, and competence with the latter taking center stage when the bundle theory is applied to the case of AI systems in the commercial context as we have seen.

Indeed, function and purpose seem to commingle in the rights theory and theory of personhood registers. A possibility would be to think of function as a binder between the more abstract "purpose" and the concreteness of "competence." It could, thus, serve as an intermediary, negotiating the proper shape of personhood between what AI systems can and should do and what we can and should make them do. This functionalist approach is, therefore, not task oriented *per se* but shifts the focus from the technical capabilities that AI systems are designed to have to the things that they are made to do for humans. In other words, the problem is put in terms of a relational approach (Coeckelbergh, 2010; Gellers, 2021). Moreover, AI systems are communicative entities, and even in moving away from considering communication as the relevant criterion for the personification of nonhuman entities, we should still consider it as relevant to whether they should be treated more like persons (Darling, 2016; Darling, 2021), including legally, since it makes us perceive them as life-like.

Communication is certainly important from a legal perspective, not least because it is what makes possible voicing internal mental states, on the expression of which rests the foundation of legal responsibility attribution. Thus, ascribing legally binding intentions to AI systems as communicative entities has been explained *via* a systems theory generalization of the "intentional stance"[23]. Intentionality is fundamental to contracting and, combined with the pervasive objectivization tendencies in contract theory springing from technological advancement, amounts to the possibility for "software agents" to make legally effective declarations of intent (Teubner, 2018) as opposed to just being a prolongation of the creators' intention. As we have already seen, however, this fails to account for the passive aspects of legal personhood as well.

LEGAL PERSONHOOD *QUA* GRADIENT

As the bundle theory unfolds, it becomes increasingly clear how such an account of personhood as a cluster concept can be mobilized with ease in fringe cases, in which not all incidents of legal personhood are at stake. This also makes the conceptual borders of personhood rather more blurred, however, raising the issue of salient criteria and thresholds and inviting reflection on whether it might not be vulnerable to a sorites paradox critique[24]. In other words, it invites the question of what makes a bundle. Because "bundle" has unclear boundaries it seems that no single

[23] "The intentional stance should be adopted when the behavior of a system is best explained and/or predicted if we attribute beliefs and desires to that system," according to Daniel C. Dennett, *The Intentional Stance*, The MIT Press, Cambridge, Massachusetts, first published in 1987, freely available online courtesy of the Internet Archive at https://archive.org/details/intentionalstanc00dani.

[24] Hyde, Dominic and Diana Raffman, "Sorites Paradox", The Stanford Encyclopedia of Philosophy (Summer 2018 Edition), Edward N. Zalta (ed.), available at: https://plato.stanford.edu/archives/sum2018/entries/sorites-paradox/.

incident added or subtracted can make a difference between a bundle and a nonbundle, and therefore, the threshold to legal personhood seems rather arbitrary.

The bundle metaphor has connotations of artificially tying together a set of nondistinct or random items, whereas a gradient might be a metaphor more apt at capturing the quality of legal personhood as a cluster property with its extension determined based on a weighted list of neither necessary, nor sufficient criteria. This, in turn, suggests different items of the same kind (in this case rights) can be added or subtracted to end up placed differently on the gradient, much like in the case of the RGB or CMYK color models for instance.

As a gradient, legal personhood is not, therefore, only a matter of adding or subtracting from a bundle of legal incidents with a minimum threshold below which we can no longer call it a bundle, but it also takes into account the kinds of items added or subtracted so that an entity can be a legal person for some specific purposes only, as in the partial legal capacity example, in which function plays a central role. This is reminiscent of the origins of the concept of legal personhood in the mask worn by ancient Greek actors on stage and that came to represent the different roles played by a person in the many areas of life and law. Vendor, partner, accused, administrator, or reasonable person are all masks one wears, sometimes superimposed, but always molded to fit them and whatever the norms of the day demanded for their protection and participation to legal life.

A loose parallel becomes possible here with David Hume's bundle theory of personal identity and the self, according to which "the peculiarly complex unity or identity of the self should be interpreted in terms of constantly changing causal relations, more like the identity of a complex play than a simple material object"[25]. What serves as inspiration here, however, is rather the gradient theory of personal identity recently attributed[26] to Anne Finch Conway (Gordon-Roth, 2018), whose views suggest a spectrum of creatures distributed in a kind of personhood gradient in which some are more or less of a person than others. A certain threshold is envisageable, but necessary conditions for passing it are not. Only sufficient conditions might be, and research to uncover the subtleties of this view is on-going.

Regardless of whether we choose to think of personhood as a bundle or as gradient, the important premise remains that legal personhood is a complex attribute in legal theory, having been expressly characterized as "gradable" aside from it also being "discrete, discontinuous, multifaceted, and fluid" (Wojtczak, 2021). This because it can contain a variable number elements of different types—such as responsibilities, rights, competences, and so on—which can be added or taken away by a lawmaker in most cases with some notable exceptions, chiefly concerning the natural personhood of humans, who cannot be deprived of their human rights, and neither can they renounce certain subjective rights. This (conveniently) mirrors the existence of thresholds posited philosophically and is also a common point between the bundle and gradient approaches to legal personhood. Both reject that anything goes when it comes to legal personhood, the latter based on the worry that such a legal instrument, malleable in the extreme, would ultimately become meaningless and ineffective in its declared purposes of protection and participation in legal life.

CONCLUSION

AI systems are in a rather singular position. We are making them show us reality in novel ways, and they are making us reconsider the way we order it in return. No matter how we formulate our answer to the legal status of AI systems question, it must acknowledge the fact that law is artifactual. Being much more in line with what we think of as such, molds are yet another helpful metaphor. They are not mere collections of things tied together by the proverbial *vinculum juris*, but tools for creating new things altogether, extensions of composing parts with their shape, size, and color situated on gradients.

Given that the skills involved in making such tools were acquired only of late by this hybrid between *homo faber* and *homo juridicus* that *homo sapiens sapiens* seems to be, they need honing. Engineering complex concepts, such as legal personhood can be looked at as a work in progress from this perspective. Applying them to such uncanny novel entities as AI systems requires the use of every other available tool in the analytical toolbox to fashion a smooth transition in the face of the overwhelming changes brought about by the advent of AI.

The underlying assumption being that AI systems' legal status is a matter of utmost importance because it determines which law is applicable and enforceable as to their uses and the ensuing consequences of those uses, this article proceeds to deconstruct that assumption by first looking at why there is a status question concerning these entities in the first place. It then inventories the possible answers to that question according to the currently entrenched legal theoretical framework and makes the case for legal creativity when it comes to the options available as to status ascription to better fit the uncanny entities that AI systems are. It then looks at methods for so doing and details one particular recent approach to solving the problem by reconceptualizing legal personhood as a bundle, which is the state of the art in our theoretical understanding. Through this new lens, it goes on to analyze "partial legal capacity" recently proposed as a solution to AI systems' legal status question. It concludes that accepting it means accepting the bundle theory of legal personhood or, at the very least, accepting that legal personhood is a cluster concept. Finally, it suggests, upon further analysis, that framing it in terms of gradient might be better suited to explain at least some use cases, AI systems included. It, therefore, sketches some incipient ideas on what could, with further research, perhaps develop into a gradient theory of legal personhood.

[25] According to the entry on "the bundle theory of the self" of David Hume, *A Treatise of Human Nature*, i. iv. 6, from Ted Honderich (ed.), *The Oxford Companion to Philosophy*, second edition, Oxford University Press, 2005.

[26] Alex Jensen, *Conway and Locke on Personhood*, detailing Heather Johnson's research as part of a project to diversify the cannon under way at the University of Minnesota, available at https://cla.umn.edu/philosophy/news-events/story/conway-and-locke-personhood.

These theoretical adjustments are necessary and significant for a more coherent answer to the legal status question of AI systems. Such an answer, well-grounded in legal theory, has the potential to influence the future legal treatment of AI systems. It can also help judges decide the hard cases involving AI systems with which they will undoubtedly be faced, not to mention help lawyers argue such cases. Perhaps most importantly, however, if such a theory succeeds in painting a clearer picture of all the relevant facets of the legal issues at stake, it could contribute to better balancing the interests of all those involved.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

The author, boursier FRESH, would like to thank prof. Christophe Lazaro for all the care, time and effort that he so selflessly dedicated to guiding the research and writing of the present article, prof. Anne-Lise Sibony and prof. Charles Pence for their insightful comments and overall support, as well as the Université Catholique de Louvain and the Institute for Interdisciplinary Research in Legal Sciences for the research opportunity and their generous financial support. With the financial support of the Fonds Spéciaux de Recherche of the French Community of Belgium.

REFERENCES

Abbott, R., and Sarch, A. F. (2019). *Punishing Artificial Intelligence: Legal Fiction or Science Fiction*. UC Davis Law Review. doi:10.2139/ssrn.3327485

Bayern, S. (2021). *Autonomous Organizations*. Cambridge University Press.

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: the Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25, 273–291. doi:10.1007/s10506-017-9214-9

Chalmers, D. J. (2020). *What Is Conceptual Engineering and what Should it Be?* Inquiry.

Chopra, S., and White, L. F. (2011). *A Legal Theory for Autonomous Artificial Agents*. Ann Arbor: University of Michigan Press.

Clark, A. (2017). "Embodied, Situated, and Distributed Cognition," in *A Companion to Cognitive Science*. Editors W. Bechtel, and G. Graham (Wiley). doi:10.1002/9781405164535.ch39

Coeckelbergh, M. (2010). Robot Rights? towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12, 209–221. doi:10.1007/s10676-010-9235-5

Crawford, K. (2021). *Atlas of AI*. Yale University Press.

Darling, K. (2021). *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. New York: Henry Holt and Company.

Darling, K. (2016). "Extending Legal Rights to Social Robots, SSRN Journal," in We Robot Conference 2012. Editors C. Froomkin, K. eds., R. Law, and E. Elgar (Cheltenham, UK; Northampton, MA, USA: University of Miami). doi:10.2139/ssrn.2044797

Dyschkant, A. (2015). *Legal Personhood: How We Are Getting it Wrong*. University of Illinois Law Review, 2075–2110.

Gellers, J. C. (2021). *Rights for robots. Artificial Intelligence, Animal and Environmental Law*. New York: Routledge

Gordon-Roth, J. (2018). What Kind of Monist Is Anne Finch Conway? *J. Am. Philos. Assoc.* 4 (3), 280–297. doi:10.1017/apa.2018.24

Gunkel, D., and Wales, J. J. (2021). Debate: What Is Personhood in the Age of AI? *AI Soc.* 36, 473–486. doi:10.1007/s00146-020-01129-1

Hermitte, M-A. (1999). *Le Droit Est Un Autre Monde*. Enquête [En ligne]. Available at: http://journals.openedition.org/enquete/1553 (Accessed July 15, 2013).

Hondius, F. W. (1980). Data Law in Europe. *Stanford J. Intern. Law* 16, 87–112.

Jowitt, J. (2021). Assessing Contemporary Legislative Proposals for Their Compatibility with a Natural Law Case for AI Legal Personhood. *AI Soc.* 36, 499–508. doi:10.1007/s00146-020-00979-z

Kiršienė, J., Gruodytė, E., and Amilevičius, D. (2021). From Computerised Thing to Digital Being: mission (Im)possible? *AI Soc.* 36, 547–560. doi:10.1007/s00146-020-01051-6

Kruger, J. (2021). "Nature, Culture, AI and the Common Good – Considering AI's Place in Bruno Latour's Politics of Nature," in Artificial Intelligence Research, First Southern African Conference for AI Research, SACAIR 2020, Muldersdrift, South Africa, 22–26, Proceedings. Editor A. Gerber (Springer), 21.

Kurki, V. A. J. (2019). *A Theory Of Legal Personhood*. Oxford: Oxford University Press

Latour, B. (2004). *Politics of Nature – How to Bring the Sciences into Democracy*. Cambridge: Harvard University Press.

Laukyte, M. (2021). *The Intelligent Machine: A New Metaphor through Which to Understand Both Corporations and AI. AI & Soc.* 36, 445–456. doi:10.1007/s00146-020-01018-7

McPherson, T., and Plunkett, D. (2020). "Conceptual Ethics and the Methodology of Normative Inquiry," in *Conceptual Engineering and Conceptual Ethics*. Editors A. Burgess, H. Cappelen, and D. Plunkett (Oxford University Press). doi:10.1093/oso/9780198801856.003.0014

Menary, R. (2007). *Cognitive Integration*. UK: Palgrave Macmillan.

Michalczak, R. (2017). "Animals' Race against the Machines," in *Legal Personhood: Animals, Artificial Intelligence and the Unborn* (Springer, Law and Philosophy Library), 91–101. doi:10.1007/978-3-319-53462-6_6

Norton, D. F., and Norton, M. J. (Editors) (2007). *David Hume, A Treatise of Human Nature: A Critical Edition* (Oxford, Clarendon Press), 27.

Novelli, C., Bongiovanni, G., and Sartor, G. (2021). A Conceptual Framework For Legal Personality And Its Application To AI. *Jurisprudence.* doi:10.1080/20403313.2021.2010936

Papakonstantinou, V., and de Hert, P. (2020). Refusing to Award Legal Personality to AI: Why the European Parliament Got it Wrong, European Law Blog. Available at: https://europeanlawblog.eu/2020/11/25/refusing-to-award-legal-personality-to-ai-why-the-european-parliament-got-it-wrong/.

Pietrzykowski, T. (2018). *Personhood beyond Humanism - Animals, Chimeras, Autonomous Agents and the Law*. Cham: Springer.

Pietrzykowski, T. (2017). "The Idea of Non-personal Subjects of Law," in *Legal Personhood: Animals, Artificial Intelligence and the Unborn* (Springer, Law and Philosophy Library). doi:10.1007/978-3-319-53462-6_4

Pottage, A., and Mundy, M. (2004). *Law, Anthropology, and the Constitution of the Social, Making Persons and Things*. Cambridge University Press.

Pound, R. (1922). *An Introduction to the Philosophy of Law*. Yale University Press, 16.

Regad, C., and Riot, C. (2020). *La personnalité juridique de l'animal (II) : Les animaux liés à un fonds (de rente, de divertissement, d'expérimentation)*. Toulon: LexisNexis.

Regad, C., and Riot, C. (2018). *Sylvie Schmitt, La personnalité juridique de l'animal (I) : L'animal de compagnie*. Toulon: LexisNexis.

Rizoiu, R. (2020). "În Spatele Oglinzii: Voința Ca Putere," in *Dreptul romanesc la 100 de ani de la Marea Unire*. Editors P. Pop, and R. Rizoiu (București: Editura Hamangiu), 579.

Schirmer, J-E. (2020). "Artificial Intelligence and Legal Personality: Introducing "Teilrechtsfähigkeit": A Partial Legal Status Made in Germany," in *Regulating Artificial Intelligence.* Editors T. Wischmeyer, and T. Rademacher (Springer), 124.

Solum, L. B. (1992). Legal Personhood for Artificial Intelligences. *North Carolina L. Rev.* 70, 1231–1288.

Sousa Antunes, H. (2020). Civil Liability Applicable to Artificial Intelligence: a Preliminary Critique of the European Parliament Resolution of 2020. Available at: https://ssrn.com/abstract=3743242.

Teubner, G. (2018). *Digital Personhood? the Status of Autonomous Software Agents in Private Law, Translated by Jacob Watson*. Ancilla Juris, 35–78.

Tuori, K. (2002). *Critical Legal Positivism*. London: Ashgate, 186–188.

V. A. J. Kurki, and T. Pietrzykowski (Editors) (2017). *Legal Personhood: Animals, Artificial Intelligence And the Unborn* (Springer), viii.

Veale, M., and Zuiderveen, B. F. (2021). Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach. *Comp. L. Rev. Int.* 22 (4), 97–112. doi:10.9785/cri-2021-220402

Wein, L. E. (1992). The Responsibility of Intelligent Artifacts: Toward an Automation Jurisprudence. *Harv. J. L. Tech.* 6, 103–154.

Wojtczak, S. (2021). Endowing Artificial Intelligence with Legal Subjectivity. *AI & Soc.* doi:10.1007/s00146-021-01147-7

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 05 April 2022
doi: 10.3389/frobt.2022.842213

Robots and AI as Legal Subjects? Disentangling the Ontological and Functional Perspective

Andrea Bertolini[1]* *and Francesca Episcopo*[2]

[1]*Scuola Superiore Sant'Anna, Dirpolis Institute, Pisa, Italy,* [2]*Università di Pisa, Department of Private Law and Scuola Superiore Sant'Anna, Dirpolis Institute, Pisa, Italy*

Robotics and AI-based applications (RAI) are often said to be so technologically advanced that they should be held responsible for their actions, instead of the human who designs or operates them. The paper aims to prove that this thesis ("the exceptionalist claim")—as it stands—is both theoretically incorrect and practically inadequate. Indeed, the paper argues that such claim is based on a series of misunderstanding over the very notion and functions of "legal responsibility", which it then seeks to clarify by developing and interdisciplinary conceptual taxonomy. In doing so, it aims to set the premises for a more constructive debate over the feasibility of granting legal standing to robotic application. After a short Introduction setting the stage of the debate, the paper addresses the ontological claim, distinguishing the philosophical from the legal debate on the notion of i) subjectivity and ii) agency, with their respective implications. The analysis allows us to conclude that the attribution of legal subjectivity and agency are purely fictional and technical solutions to facilitate legal interactions, and is not dependent upon the intrinsic nature of the RAI. A similar structure is maintained with respect to the notion of responsibility, addressed first in a philosophical and then legal perspective, to demonstrate how the latter is often utilized to both pursue ex ante deterrence and ex post compensation. The focus on the second objective allows us to bridge the analysis towards functional (law and economics based) considerations, to discuss how even the attribution of legal personhood may be conceived as an attempt to simplify certain legal interactions and relations. Within such a framework, the discussion whether to attribute legal subjectivity to the machine needs to be kept entirely within the legal domain, and grounded on technical (legal) considerations, to be argued on a functional, bottom-up analysis of specific classes of RAI. That does not entail the attribution of animacy or the ascription of a moral status to the entity itself.

Keywords: legal subjects, personhood, agency, responsibility, autonomy, liability, electronic personhood, risk-management

Edited by:
David Gunkel,
Northern Illinois University,
United States

Reviewed by:
Simon Chesterman,
National University of Singapore,
Singapore
Kamil Mamak,
Jagiellonian University, Poland

Correspondence:
Andrea Bertolini
andrea.bertolini@santannapisa.it

Specialty section:
This article was submitted to
Ethics in Robotics and Artificial
Intelligence,
a section of the journal
Frontiers in Robotics and AI

Received: *23 December 2021*
Accepted: *11 March 2022*
Published: *05 April 2022*

Citation:
Bertolini A and Episcopo F (2022)
Robots and AI as Legal Subjects?
Disentangling the Ontological and
Functional Perspective.
Front. Robot. AI 9:842213.
doi: 10.3389/frobt.2022.842213

INTRODUCTION

Whether advanced robots and AI applications (henceforth, RAI) are, should, and eventually will be considered as "subjects" rather than mere "objects" is a question that has strongly characterized the social, philosophical, and legal debate since Solum's seminar article on "Legal Personhood for Artificial Intelligence" (Solum, 1992), and arguably even earlier (Turing, 1950; Putman, 1964; Nagel, 1974; Bunge, 1977; Taylor, 1977; Searle, 1980; Searle, 1984; McNally and Inayatullah, 1988). However, debates have significantly intensified over the last two decades, with interest in both the scientific and non-academic circles raising every time a new technology rolls out (e.g., autonomous cars being tested in real-life scenarios on our streets), or an outstanding socio-legal development occurs (e.g., the humanoid Sophia receiving Saudi Arabian citizenship)[1] (see, e.g., Allen et al., 2000; Allen et al., 2005; Teubner, 2006; Chrisley, 2008; Coeckelbergh, 2010; Koops et al., 2010; Gunkel, 2012; Basl, 2014; Balkin, 2015a; Iannì and Monterossi, 2017; Christman, 2018; Gunkel, 2018; Nyholm, 2018; Pagallo, 2018b; Santoni de Sio and van den Hoven, 2018; Lior, 2019; Loh, 2019; Turner, 2019; Wagner, 2019; Andreotta, 2021; Basl et al., 2020; Bennett and Daly, 2020; Dignum, 2020; Gunkel, 2020; Kingwell, 2020; Osborne, 2020; Powell, 2020; Serafimova, 2020; Wheeler, 2020; De Pagter, 2021; Gabriel, 2021; Gogoshin, 2021; Gordon, 2021; Gunkel and Wales, 2021; Joshua, 2021; Kiršienė et al., 2021; Martínez and Winter 2021; Schröder, 2021; Singer, 2021).

In the policymaking arena, a recommendation from the European Parliament famously urged the European Commission to consider whether robots could be attributed an "electronic personality" (European Parliament, 2017), but the idea didn't gain momentum and found no place in the most recent initiatives on the regulation of RAI, some of which seem to dismiss the possibility in a surprisingly sweeping fashion (European Commission, 2018; European Parliament, 2020). Yet, with social robots soon to be incorporated into our lives, a sound discussion of whether—to borrow the Editors' own words—"robots, AI, or other socially interactive, autonomous systems have [or will ever have] some claim to moral and legal standing"[2] becomes inescapable.

Engaging with some of the most prominent literature in the field, the paper seeks to answer the second prong of this question, i.e., whether robots, AI, or socially interactive, autonomous system have some claim to *legal* standing.

The contribution that the paper seeks to make is threefold.

First, the paper develops a specific framework to disentangle the conceptual and analytical knots, whose obfuscating presence often misleads even the most insightful analyses of the matter. The framework is based on three major distinctions, which the vast and heterogenous debate on RAI's standing needs to acknowledge and take into consideration: i) between the legal and the moral domain, and between the respective notions of "responsible subject"; ii) between the fully fledged and the limited notions of subjectivity; iii) between the ontological/essentialist and the functionalist/consequentialist grounds of standing.

Secondly, the paper discusses some fundamental concepts which come into play in the discussion of moral and legal standing of RAI—i.e., those of agency, responsibility, and personality—, to lay the ground for a shared understanding of the debate.

Thirdly, and applying the methodological and conceptual tools described above, the paper argues that: i) at the current stage, there are no ontological reasons why RAI need to be considered legal subjects; and ii) there may nevertheless be functional reasons to do so in particular cases, when endowing them with specific rights and obligations proves the best way of fostering the individual and social interests that the law is meant to protect.

Against this backdrop, the paper is structured as follows.

In §2 we introduce some of the traditional claims for treating RAI as subjects and identify a series of conceptual and analytical problems. Moving from these considerations, we sketch the analytical framework shape distinguish the various perspectives, which we believe that a sound a coherent discussion of RAI's standing should follow.

In §3, we put said analytical framework into practice. We first narrow down the ultimate scope of the inquiry, relating to a generalized moral and legal standing, but rather to RAI's specific capacity to qualify as subjects legally accountable for the illegal or wrongful actions and events caused. Accordingly, we disentangle the legal from the moral dimension of standing and move on to consider when an entity may be granted a particular legal qualification and be subjected to a given legal regime, separating what we refer to as, respectively, the ontological and the functional viewpoints.

Following this line of argumentation, §4 proves that, at this stage, there are no ontological reasons to consider RAI as legal subjects. §5 then adopts the functional perspective and argues that, despite ontologically qualifying as objects and not subjects, it may nevertheless be appropriate and desirable, under certain circumstances, to grant specific technological applications with limited and narrow forms of legal personality.

In conclusion, §6 sums up the main arguments and uses them to critically discuss the European Parliament proposal of October 2020 (European Parliament, 2020), which seems to categorically exclude the possibility to treat RAI as legal subjects.

TOURING THE RAI'S SUBJECTIVITY FOREST: AN ANALYTICAL FRAMEWORK

The literature on RAI's subjectivity is vast and varied. However, two threads seem particularly relevant.

On the one hand, it has been claimed that a robot, an intelligent artefact, or other socially interactive mechanisms, may be due to some level of social standing or respect. That

[1] In the generalist press, see, respectively: https://www.reuters.com/technology/google-self-driving-spinoff-waymo-begins-testing-with-public-san-francisco-2021-08-24/, https://www.theverge.com/2020/12/9/22165597/cruise-driverless-test-san-francisco-self-driving-level-4; https://www.nationalgeographic.com/photography/article/sophia-robot-artificial-intelligence-science; https://www.businessinsider.com/meet-the-first-robot-citizen-sophia-animatronic-humanoid-2017-10, https://www.forbes.com/sites/zarastone/2017/11/07/everything-you-need-to-know-about-sophia-the-worlds-first-robot-citizen/?sh=475456c46fa1 (all articles last accessed on 15 December 2021). For a recent review of the debate, see Schröder, W.M. (2021). "Robots and Rights: Reviewing Recent Positions in Legal Philosophy and Ethics," in *Robotics, AI, and Humanity: Science, Ethics, and Policy,* eds. J. Von Braun, M. S. Archer, G.M. Reichberg and M. Sánchez Sorondo (Cham: Springer International Publishing), 191–203.

[2] Research topic description, available at https://www.frontiersin.org/research-topics/17908/should-robots-have-standing-the-moral-and-legal-status-of-social-robots (last accessed 15 December 2021).

they seem to have the "psychological capacities that we had previously thought were reserved for complex biological organism such as humans" (Prescott, 2017). That they are "worthy of moral value," if not moral subjects *tout court*, and that not giving them legal standing would constitute a violation of their rights, as well as an impoverishment of our ethical stance as human beings. In these terms, the fight for RAI's rights is frequently framed as another step in the corrective evolution of our legal systems, which has progressively expanded the legal recognition of previously discriminated humans, and is now opening towards non-human entities—animals, rivers, idols, etc.—(Gunkel, 2018; Kurki, 2019; Gellers, 2021).

At the same time, it has been claimed that some RAI are so technologically advanced, that they invite "a systemic change to laws or legal institutions in order to preserve or rebalance established values" (Calo, 2015, 553). In this sense, they should be recognized as subjects, having rights and duties of their own comparable, if not identical, to those of natural persons (Floridi and Sanders, 2004; Matthias, 2004; Stahl, 2006; Teubner, 2006; Matthias, 2008; Koops et al., 2010; Matthias, 2010; Gunkel, 2012; Floridi, 2014; Calo, 2015; Schwitzgebel and Garza, 2015; Richards and Smart, 2016; Gunkel, 2018; Nyholm, 2018; Danaher, 2020; Gunkel, 2020; Nyholm, 2020; Gunkel and Wales, 2021). In particular, it is often argued that being their actions so much outside humans' control, we should deem them responsible for the wrong caused, instead of blaming the producer, the owner, or the user behind them (Matthias, 2004; Stahl, 2006; Matthias, 2008; Purves et al., 2015; De Jong, 2020; Gunkel, 2020).

Having a comprehensive view of the debate is particularly important, as it shows the plurality of concerns at stake in the discussion on the status of RAI in our society, whose viewpoints and analytical tools overlap and complement one another only in part. With a certain degree of approximation, there are three orthogonal strands of analysis worth identifying.

First, the current debate on the subjectivity of RAI sits at the crossroad of different disciplines: engineering, computer science, law, philosophy, sociology, to name but a few. If cross-fertilization and plurality of perspectives are to be fully welcomed, some caveats are needed to avoid negative side-effects.

Secondly, in social sciences, the debate on the subjectivity of RAI is shaped around a series of overlapping concerns and questions. In our opinion, the most important distinction to be drawn is the one between those who discuss what moral and legal entitlements RAI may and possibly should be granted as the main research question, and those who come to it indirectly, as part of an inquiry which has the focus somewhere else—in the example above, the allocation of responsibility for illegal or wrongful actions and events. The very interest and sensibility between the two viewpoints differ radically: asking the broad theoretical question of whether robots have a claim to moral and legal entitlements not only is broader in its scope and implications, but often responds to a peculiar "robot-centered" approach: despite considering both positive and negative entitlements the starting point is commonly the recognition of robot's *rights* for the robot's own sake, or at least for a coherent and correct explication of the moral or legal system. Conversely, those who discuss the issue of robotics' entitlements indirectly, as a means of addressing specific problems, often have a "human-centered" standpoint: raising robots to the status of subjects is commonly presented as a way of solving what we, as humans, consider a moral or legal problem.

Thirdly, and finally, the debate on the subjectivity of RAI may be distinguished based on the grounds according to which the latter are considered as worthy (or unworthy) of raising to the status of subjects. We identified two major approaches, which we call, respectively, "ontological" or "essentialist," and "functional' or "consequentialist". The first answers the question of RAI's standing moving from the properties they display, while the other bases its answer on the consequences which derive from their legal qualification.

Taken together, these distinctions are of fundamental importance. Not only do they work as critical tools, allowing us to dissect and fully understand the state of the discussion—what claims exactly are made, for which purposes' and upon which grounds. They also constitute essential tools for constructing a sound analytical framework, which could help us address the question of electronic personhood.

Differentiating law and morality teaches us two lessons. One is conceptual: we need to avoid the temptation to automatically translate assumptions or standards pertaining, e.g., to moral philosophy, and elevate it as a ground for legal reform (Fossa, 2021). The second is broader and relates to the relationships between the different domains at stake. Despite the important interactions between philosophical and legal analysis, whether RAI should gain something akin to an electronic personhood is only partially dependent on the moral status of such technologies and should thus be discussed in a proper legal perspective. While the legal and philosophical approaches find some points of convergence in the discussion on what properties would make RAI "moral agents," they diverge whenever the focus is on whether, and if so how, attributing them legal entitlements would foster the ends of the legal system.

The distinction among different research questions forces us to be analytically clear and coherent. On the one hand, it teaches us against conflating the issue of whether robots may be bearer of rights and duties in general with that of their responsibility and accountability. On the other, it forces us to choose, among the various manifestation of "standing," what exactly to address in the substantive part of the inquiry, precisely to avoid unpreceded claims, whose province and implications would be hard to tame.

Disentangling the two possible approaches according to which something may or should qualify as a subject is equally fundamental. First, before arguing for a solution based upon a specific approach, it is important to question whether the latter is accepted by the system under analysis—moral or, in our case, legal—and, if so, which role it may legitimately play. Secondly, arguing an identical conclusion in terms of policy recommendation bears radically different theoretical and practical consequences, depending on which of the two perspectives is adopted. If we say that robots should be held responsible because they are the "subjects"—and not a mere tool in the hands of a human—thence not only their liability may follow but also complex bundles of rights and obligations intended to protect their own interest. Even if the original question only concerns RAI's standing as per their accountability, the solution would impact their overall legal status. The discussion on rights and duties on the one side, and liability on the other, would ultimately converge. Instead, if RAI are treated as juridical persons with the sole aim of segregating selected assets (shielding human beings from the legal and economic consequences of its operations, and eventually providing a diversified taxation scheme), then the overall legal—and ethical—implications radically differ (Bertolini, 2013; Pagallo, 2018a). The two stances must not be confused.

We will now address the question—are or should RAI qualify as subjects legally accountable for their own actions—following the various steps introduced above.

LAW, MORALITY, AND THE GROUNDS OF LEGAL SUBJECTIVITY

Law and morality are two normative systems that control and regulate social behaviors, which may be framed as at least partly independent. In a legal system, positive and enforceable standards of conduct guide a community, preventing conflicts and incentivizing desirable behaviors, and offer second-order criteria for identifying, modifying, and enforcing said rules (Marmor and Sarch, 2019).[3] On the contrary, morality comprises of those principles that society deems relevant for distinguishing between right and wrong (Gert and Gert, 2020), and offers a code of conduct valid irrespective of what is legally enacted. The ultimate relationship between the two domains constitutes one of the oldest and major questions in jurisprudential studies (Bentham, 1823; Austin et al., 1954; Dworkin, 1977; Dworkin, 1986; Wren et al., 1990; Coleman, 2003; Kelsen et al., 2005; Raz, 2009; Finnis, 2011; Hart, 2012; Hershovitz, 2015; Dickson, 2021). Positivist theories hold that legal normativity is autonomous and distinct from that of morality, and its validity is not dependent on its content (Green and Adams, 2019). Naturalist theories, on the contrary, argue that law and morality are interdependent and that to regulate social behaviors, the law must have moral content (Finnis, 2020).

Yet—even the most extreme of the naturalist theorists would concede—the moral relevance of a matter cannot be considered *per se* the source of legal normativity and deriving one from the other would be a serious mistake. This does not mean that moral considerations have no role in the legal domain, but rather that they must be contextualized within the space attributed to them by the legal system. If we want to discuss the legal status of RAI—the starting point needs to be the grounds upon which a given order qualifies entities, for the sake of regulation (Bryson et al., 2017). The question then becomes: how does a legal system "decides" how to qualify different entities? (Kurki, 2019).

Regulation—and legal reforms in particular—may be grounded in two different approaches (Bertolini, 2013).

According to the "ontological" or "essentialist" perspective, entities have a clear-cut legal qualification based on their inherent features, which in turn determines the applicable legal rules. Pursuant to such a narrative, we may need to adopt new rules, or change existing ones, when the object of the regulation (in this case, RAI applications) is so different from what we have been regulating so far (other, less advanced forms of technology), that a distinct legal qualification is due. In the current debate, such a stance claims the necessity to elaborate an alternative and potentially intermediate category between that of subjects and objects of law (Calo, 2015). The first notion encompasses those that within the legal system are attributed rights, the latter those entities upon which rights insist, and are exerted by the former. However, so defined, the duality of the alternative appears logically necessitated, to the point that an intermediate category would be altogether inconceivable and useless in a technical legal perspective. Indeed, either one entity is solely capable of being subject to someone's rights—hence it's an object –, or is able to possess rights. *Tertium non datur*. The circumstance that the law treats some entities such as corporations possessing rights for the sake of a given legal relation, while, in others, considers them as the objects upon which rights are exerted, simply means that the distinction between subject and object may be contingent upon different legally relevant circumstances, and does not lead to the existence of an intermediate category (Kurki, 2019).

If this is true from the ontological perspective, the functional one has quite a different approach. Indeed, the latter claims that legal frameworks shall be developed according to their adequacy in performing the functions attributed to them, as well as the broader consequences deriving therefrom. In this view, a particular legal qualification, and the rules applicable thereto, should be adopted based on the desirability of social, legal, and economic implications they bring about (Bertolini, 2013; Bertolini, 2014; Balkin, 2015a; Palmerini and Bertolini, 2016; Bryson et al., 2017).

Our legal systems commonly work on a combination of the two approaches: there are specific features that justify the qualification of an entity as a legal subject and, in addition, *ad hoc subjectivity* is sometimes granted for functional reasons. Regardless of whether these represent "legal fictions," or mere expression of the legal system's normative power to recognize

[3] A simple yet effective overview of this view can be found under the entry "Legal Positivism," offered in the Oxford Dictionary of Law (John Law (ed), Oxford: Oxford University Press (2018)): "An approach to law that rejects natural law and contends that the law as laid down (*positum*) should be kept separate—for the purpose of study and analysis—from the law as it ought morally to be. In other words, a clear distinction must be drawn between "ought" (that which is morally desirable) and "is" (that which actually exists). The theory is associated especially with the thought of Jeremy Bentham (1748–1832), John Austin (1790–1859), H. L. A. Hart (1907–1992), and Hans Kelsen (1881–1973), who differ from one another in important respects but generally adhere to the above separability thesis. In addition, legal positivists normally adopt the so-called social fact thesis (that legal validity is a function of pedigree or related social facts) and the conventionality thesis (that social facts giving rise to legal validity are authoritative by virtue of social convention)".

rights and duties, what is important is that the two approaches may very well coexist.

The following paragraphs further elaborate on this point, disentangling and critically evaluating the various arguments underlying the call for "artificial personhood" under both the ontological and the functional perspectives.

RAI AS *SUBJECTS*? THE ONTOLOGICAL PERSPECTIVE

Both the idea that we shall avoid the so-called "responsibility gap"—where humans are forced to compensate damages for which they have no or very limited control, and that machines shall behave as responsibly as possible, according to the principles elaborated through "machine ethics" (Wallach and Allen, 2009a; Anderson and Anderson, 2011) –, and that according to which RAI may have rights of their own, are often expressly grounded on the belief that the peculiar features displayed by advanced RAI (their asserted autonomy and ability to modify themselves) make them agents; more specifically, moral and possibly legal subjects, who should consequently be held responsible for their actions (Allen et al., 2005; Wallach and Allen, 2009b; Howard and Muntean, 2017).

However, the ontological claim according to which RAI's essential qualities make them subjects, rather than mere objects, is far from being proved (Fossa, 2018).

This consideration, begs for a further question. Indeed, if we are to define what a robot can and cannot do by referring to the notions of subjectivity or personhood, agency, responsibility or liability, it is first necessary to understand what we mean by these concepts, which have complex and possibly indeterminate meanings.

As anticipated above, when discussing the challenges and opportunities brought about by RAI, both economic, legal, ethical, philosophical, and engineering considerations come into play, leading the debate to merge the methodological and analytical background of heterogeneous disciplines. Yet, economists, engineers, philosophers, and lawyers may use terms that have both a common, a-technical understanding and one which is peculiar of their own subject. Therefore, engineers or lawyers may speak of autonomy to denote different qualities than the ones that philosophers understand as associated with the said notion (Haselager, 2005). This constitutes a case of semantic ambiguity. Both the meaning of a concept and the conditions of its use depend on the context in which the latter is used, so that the transmission of a notion from one context to the other represents a process of "semantic extension," which may lead to substantial confusion (Waldron, 1994; Endicott, 2000).

As highlighted by the studies on legal reasoning and linguistic indeterminacy (Waldron, 1994; Endicott, 2000), unclear and under-specified terminology may compromise the acceptability of the warranties used to back a specific argument, which in turn affects the correctness of the overall claim (Toulmin, 1964; Alexy, 1978).

The Philosophical Notion(s) of Subjectivity and Agency

Trying to identify and condense the philosophical debate on what is a "subject" is a dauntingly difficult task and one which is not our intention to embark on. In essential terms, we may define a subject as an entity that relates with another entity that exists outside itself—the object— through a relationship which the subject enters by means of personal experience and/or consciousness (Thiel, 2011).

In continental philosophy, the discussion on "subjectivity" strongly relates to that of "agency" and "moral status". In this section, we will consider the former, while § 4.3 will discuss the latter.

From a philosophical perspective, agents are subjects who can act—i.e., perform actions—while agency denotes the manifestation of such capacity (Schlosser, 2015). However, "actions are doings, but not every doing is an action" (Himma, 2009): according to the main variations of the "standard conception", an event may be deemed as an action only if brought about intentionally (Anscombe, 1957; Davidson, 1963) thus not being the mere result of causal determinations among naturalistic events.

In turn, intentionality is often defined as "the determination of a specified end that implies the necessity of actions of a specified kind" (Gutman et al., 2012).

According to some authors, the kind of rationality required for intentional performance consists in being capable of rationally justifying one's actions in reference to determined and determinable purposes, which, in turn, requires the deliberative and argumentative skills that only human beings possess, let alone because of their linguistic abilities. Under this view, only humans can perform actions, being able to reason and decide intentionally (Frankfurt, 1971; Taylor, 1977; Gutman et al., 2012).

Other theories set a lower threshold, describing intentionality as a mental state—such as belief, desire, will—that does not necessarily entail human-like rationality, and rather extends to the spontaneous initiation of actions that do not follow rationally justifiable desires (Ginet, 1990). Pursuant to this idea, "X is an agent if and only if X can instantiate intentional mental states capable of directly causing performance" (Himma, 2009).

However, this begs the question of how to detect mental states, whether they are non-physical subjective experiences or rather objective attitudes in the physical structure of the entity. Even if the very essence of mental states is difficult to grasp, some still read them as requiring a certain capacity of introspection, and thus of consciousness—but how to determine its existence, or set the relevant threshold required, is uncertain (Himma, 2009). Against this "hard problem", some suggested to presuppose consciousness, unless proved otherwise, and treat an entity as having such capacity based on the performative equivalence of their doings with those of beings whose consciousness is not contested (Dennett, 1991; Frankish, 2016; Dennett, 2018).

In opposite terms, some authors have theorized a "minimal agency" which contests the need for "mental states" and qualifies as agent any unified entity that is distinguishable from its environment and that is doing something by itself according to certain goals. Pursuant to this view, very simple organisms can be said to have the intrinsic goal of continuing their existence, even if they lack the ability to rationally elaborate and justify their aims and actions (Barandiaran et al., 2009; Gunkel, 2018, pp 96-105).

The discussion of the qualification of RAI as agents is strongly debated, and it would fall beyond the scope of the paper (as well as the capacity of the authors) to solve it once and for all.

Nevertheless, from the above discussion, we can derive an important insight: the definition of agency constitutes a more basic notion than other compound concepts, such as those of rational, conscious, introspective, autonomous agency and the like (Himma, 2009). While it is possible to consider an agent as a "subject," it is debatable that a mere agent—so loosely defined, without reference to rationality, consciousness, and intentionality—would meet the threshold relevant for legal consideration in an ontological perspective.[4]

As we will see in the following sections, this specification is of crucial importance also because, despite the variety of discourses which are made on the topic, the statement that RAI applications should qualify as agents—and thus be held morally and legally responsible—is based precisely on the (not always explicit) assumption that they are not mere agents, but rather *autonomous agents*.

Indeed, the idea of intentionality certainly goes towards (without necessarily overlapping) that of autonomy. Margaret Boden famously claimed that: "[a]n entity is autonomous when its behaviour-directing mechanisms may be shaped by the entity's experiential history, are emergent in nature, and are reflectively modifiable by that entity", deriving from this that "an individual's autonomy is the greater, the more its behavior is directed by self-generated (and idiosyncratic) inner mechanisms, nicely responsive to the specific problem-situation, yet reflexively modifiable by wider concerns" (Boden, 1996). In similar terms, Gutman and colleagues define an autonomous entity as one whose actions are i) free, in the sense of resulting not from external coercion but rather from one's own deliberation and ii) are means to achieve ends which are set by the subject himself (Gutman et al., 2012). Condition i) sets the standards that we have already discussed, namely, that an action is to be contrasted to a mere behavior, a deterministically caused event that was not brought about intentionally. What differentiates the notions of intentionality and autonomy is that the latter puts major importance on the origin of the goals for which the actions are performed. Defining an entity as an autonomous agent—instead of a mere agent—implies that the former has acted to obtain its own goals.

The Legal Notion(s) of Subjectivity and Agency

As a social construct, the definition and attribution of legal personality is subject to historical and cultural changes. Indeed, twenty-first century developments—such as the raise of environmentalist and animalist concern, as well as artificial intelligence, and corporate personhood—compelled us to critically consider who, or what, is a "person" according to the law, and how our understanding of legal personhood came about (Kurki, 2019).

In the modern western legal tradition, the "orthodox view" (Kurki, 2019) sees legal subjectivity or personhood as the capacity to hold legal positions, such as rights and duties.[5] Each person has said status from the moment of birth until their death, being banned forms of *capitis deminutio*, such as those related to slavery in ancient Rome or to political and racial prosecution of Jews in the Nazi regime.[6] This means that the exclusion of legal personhood to certain categories of human beings is prohibited, although foreign national or stateless person may lack the capacity to hold some rights, with the exclusion of human rights, which belong to everyone because of their human being. In a specular way, embryos and fetuses are also granted specific safeguards, and may be attributed ad-hoc legal rights—particularly some personal rights (like that to health) and patrimonial (heirship)—despite not qualifying as "natural persons".[7]

However, legal capacity is not an exclusive feature of human beings: non-human entities—such as corporations and associations—may be granted general legal capacity, thus being capable to bear those rights and duties which do not require the

[4] Indeed, functional considerations might lead to different conclusions, but there it is not the notion of agency in its philosophical dimension that matters, see §5.

[5] In recent times, the concept of legal personality has been challenged by external pressures: the limitation of "natural personhood" to human beings is allegedly harder and harder to justify, but the legalist alternative of "everything goes" is condemned as unworkable and counterproductive. Against these considerations, the very notion of legal personality is undergoing a new phase of scrutiny. Some have gone so far as to contest the correctness of the "orthodox view," suggesting that legal personality should be seen not as a gradual property, where some essential elements of a broader "bundle of personality incidents" are attached to an entity Kurki, V.a.J. (2019). *A theory of legal personhood*. Oxford, United Kingdom: Oxford University Press. These suggestions are certainly worthy of careful considerations. Yet, the critique to the notion of legal personality seems unnecessary, as it is the rejection of the binary alternative between legal subjectivity and lack thereof. While a proper discussion of the matter would fall outside the scope of this paper, it is here important to recall that two important contributions made by said renewed conception could be incorporated within the traditional—and commonly accepted—understanding of legal personality. First, the idea of legal subjectivity as a boundless of incidents can be usefully incorporated in the understating of legal persons as "entities capable of holding legal positions," in the sense that it helps clarifying the various configurations that legal personhood may have. Indeed, only humans are considered has having a fully-fledged legal personality, whereas other entities may well be recognized as subjects whenever attributed specific rights and duties, without automatically acquiring the capacity to hold other forms of entitlements. The distinction that Kurki makes between legal subjects and entities that merely qualify as rights or duties holder is arguably better framed as acknowledging different degrees or extensions of legal capacity.

[6] In Italy, for example, natural persons acquire legal capacity with birth (art. One of the Italian Civil Code), and no one can be deprived of it for political reasons (art. 22 Italian Constitution). References on this matter may only be minimal Falzea, A (1989). "voce « Capacità (teoria gen.)»," in: *Enciclopedia del diritto* (Milano: Giuffrè)., 8 ff.

[7] On the legal status of embryos and fetuses, Jost, T.S. (2002). Rights of Embryo and Foetus in Private Law. *American Journal of Comparative Law* 50., Seymour, J (2002). The legal status of the fetus: an international review. *J Law Med* **10,** 28–40.

holder to be a human being (thus excluding, e.g., those arising from marriage). Organizations set up to undertake an activity may thus qualify as "persons" and treated as autonomous and separated from the natural persons owning and administering them—although in exceptional occasions the veil of asset partitioning can be lifted, making shareholders personally liable for the debts of the corporation (Kraakman et al., 2017, 5 ff).

Thus, in the legal dimension, being an agent equals having "legal capacity," whereas a narrower version of this notion merely covers the "legal capacity to act".

Indeed, despite possessing legal personhood, legal subjects may still lack the legal capacity to act, i.e., the ability to autonomously modify one's rights and duties by performing legal acts. This constitutes a first fundamental definition of "agency" in legal terms.[8]

To be correctly understood, such notion shall be complemented with a taxonomy of legally relevant facts and acts, which—with some variations (e.g., in the legal, doctrinal, or jurisprudential formants—Sacco, 1991)—may be found in various jurisdictions belonging to the European continental legal tradition[9].

Indeed, "facts" denote naturalistically caused events or human behaviors producing specific legal effects, where—if having human origin—it is immaterial whether they were brought about intentionally or not. On the contrary, "acts" constitute intentional actions which the law considers as the basis to produce given legal effects. Among the latter, we could further distinguish among: i) "mere acts", where the action itself is intentional, but the legal effects are produced regardless of whether the author intended to bring about such legal consequences or not; ii) "juridical acts", which produce their peculiar legal effects only if the action was performed intentionally as a means to achieve specific consequences; said otherwise, the production of legal effect is not a mere by-product of the action, by rather the reason why the latter was undertaken.

What has been said so far does not mean that the actions of those who lack the legal capacity to act have no legal effect, or that they do not have the power to perform legal actions at all. On the contrary, any entity—even non-human entities—may cause events, for which the law sets specific legal consequences, despite no legal capacity being required therefor. For a person to perform mere acts, it is necessary to have what is called "natural capacity", i.e., having the ability to understand the meaning and consequences of one's own actions, and to act accordingly. For example, if an underage child, having full intellectual capacity, causes physical damage to another person with fault or malice, she would still be liable for the wrong caused (even though, under certain conditions, her parents would be called to bear the economic consequences). On the contrary, full legal capacity is required for entering a valid contract or performing other juridical acts. If we assume that the same under-age person may be a real-estate owner and wanted to sell a property, despite having the legal capacity (as far as the ability to be entitled with property rights is concerned), she would lack the power to enter a legally valid contract, and need someone else acting on her behalf, namely an agent. This leads us to another point worthy of discussion.

Indeed, in a narrower sense, the term "agency" also refers to that institution, or rather set of norms, allowing and regulating the fiduciary relationship whereby a subject—the "agent"—is expressly or implicitly authorized to act on behalf of another subject—the "principal"—to create legal relations between the latter and third parties. Thus, an agent who acts within the scope of authority conferred by his or her principal—or so long as a third party in good faith may legitimately believe her to do so—binds the principal to the obligations she creates vis-a-vis third parties. However, for such effects to be produced, it is not necessary for the agent to have legal capacity, but only for the principal.

Against this background, the relevant question then becomes whether RAI could be "legal subjects" and, if so, whether they could only cause legal fact or also legal act. As for the first issue, it seems that the alternative is either recognizing the fully fledge status comparable to that of "natural persons," if they are deemed to have essentially similar features to that of humans (and no functional reasons justifying not doing so!) or attribute them ad hoc legal personhood similarly to what we do with corporations. While the second option is, in technical terms, possible and compatible with the tools offered by the legal systems, the first one depends on our understanding of the relevant properties that would make a robot sufficiently like us, to justify its qualification as a legal subject (Kingwell, 2020; Osborne, 2020; Jowitt, 2021)—which we seek to identify throughout this paper.

For the moment, it is interesting to consider the second question addressed below, namely, whether RAI could perform legal acts. As for legal act *stricto sensu*, the question is again, whether their autonomous actions could qualify as "intentional" for the purpose of the legal system. Otherwise, it would constitute merely a legal fact. If, on the contrary, it could produce such an effect, then the behavior would qualify as a legal act and possibly a juridical act. However, from a legal perspective, this does not mean that robots would necessarily become fully-fleged subjects: their role may resemble that of the agent, who acts towards the end set by the principal, and thus produces effects within the legal sphere of the latter, being able to choose how to perform the

[8] According to our previous example (the Italian legal system), one subject acquires the capacity to act when he or she become of age—turns 18 years old—(art. Two Italian Civil Code) and can be limited or revoked by the courts, for example through interdiction, i.e., by depriving the person of the right to handle his or her own affairs because of mental incapacity (art. 414 ff. Italian Civil Code). See Stanzione, P (1988). "voce «Capacità I diritto privato»,", in: *Enciclopedia giuridica* (Bologna-Roma: Zanichelli-Foro it.).

[9] Indeed, variations on these distinctions exist between different legal systems. The tripartite structure is typical of German, law, which differentiate between juridical facts, juridical acts, and legal transaction. On the contrary, French law expressly differentiate only between juridical acts—legal transactions—and juridical facts, but the latter are thought to encompass both what we here identify a juridical facts *stricto sensu* and juridical acts *stricto sensu*, and indeed attaches different legal consequence to each category. Italian law, instead, distinguishes between "fatti giuridici" and "atti giuridici," but legal scholarship follows the German model, and it predominantly (although not unanimously) acknowledges the category of "negozi giuridici" (i.e., legal transactions) as opposed to that of "atti giuridici in senso stretto" (i.e., other legal acts). For a synthetic but effective reconstruction of this issue, see Sirena, P (2020). *Introduction to Private Law*. Il Mulino.

intended task—including, for instance, concluding contracts. Indeed, the law allows the production of effects on another subject, who is held responsible for having identified the desired results, regardless of the level of autonomous agency displayed by the entity who performed the action. Just like a person may be bound to the legal effects produced in her legal sphere by the contract signed by a representative—an adult with full legal capacity, who has the maximum autonomy in determining the content of the agreement—, she may as well be bound by the effects produced by the action of a machine—certainly showing a lower degree of autonomy than the corresponding human agent—whose activity was initiated or requested by him, and who identified the need the system was to fulfil.

RAI as *Accountable* Subjects? The Philosophical Notion(s) of Responsibility

According to the traditional philosophical discourse based on Aristotleian ethics (Aristotle, 1985), moral responsibility is the state which characterizes the subject whose actions are judged as worthy of praise or blame (Eshleman, 2016).

According to the perspective adopted, moral responsibility may be either merit-based—so that praise or blame would be an appropriate reaction toward the candidate only if s/he deserves such reactions—or consequence-based—so that moral judgment would be appropriate only when they are likely to have the desired effect in the agent's actions and dispositions –. In this paper, we will take into consideration the merit-based approach, as the major reactions to morally reprehensible actions take the form of legal sanctions (broadly intended, i.e., considering different forms of liabilities) (Bobbio, 1969; Hart, 2012). The consequence-based approach to moral responsibility, on the contrary, shall thus be reframed as a peculiar form of the functional approach to the ascription of liability, which will be considered in the following section.

In this sense, one's action may be a candidate for moral evaluation, only if she i) could exercise control over her actions and dispositions, and ii) was aware of what she was bringing about. These are generally referred to as the control and the epistemic conditions (Eshleman, 2016).

For the sake of this argument, we will leave aside the deterministic problems connected to one's ability to control her actions and dispositions, and merely assume that i) agents have a certain degree of freedom of determination and ii) the practice of holding someone responsible needs no external justification in the face of determinism, since moral responsibility is based on social intrinsic reactive attitudes (Strawson, 1962).

That being said, it is necessary to ask whether a machine could meet the control condition. Again, this question must be addressed considering the peculiar form of "autonomy" that current RAI display. Indeed, they lack what is commonly referred to as "strong autonomy," i.e., the ability to decide freely and coordinate one's action towards a chosen end, and only have a "weak autonomy," i.e., the capacity to decide, without external input or human supervision, between different possible ways of performing a given task or achieving a given goal. Even in a scenario where the machine learns from the environment, possibly adapting its functioning as a result of this interaction and learning, the machine cannot be said to be in control of its actions: even if it is free to determine the way in which to act, its choice is still determined by the need to interactively adjust its functioning to the environment and, on the basis of the available data, plan the most efficient way of performing its tasks. Given that the machine does not have control over the goals which it is programmed to achieve, since they are set by humans (most likely, the programmer), it cannot be deemed in control of the end itself (Gutman et al., 2012; Bertolini, 2013).

Likewise, artificial moral responsibility could not be recognized because it would still lack the epistemic condition. In the philosophical debate, the issue of awareness is separated by that of the possible deviancy of the causal chain initiated with one's own actions, which, if anything, shall be traced to the definition of agency, not of moral responsibility (Schlosser, 2015). Awareness is rather to be understood as "the interpretive process wherein the individual recognizes that a moral problem exists in a situation or that a moral standard or principle is relevant to some set of circumstances" (Rest, 1986). One entity's complete and unavoidable lack of moral awareness equals the impossibility of its moral consideration (Brożek and Janik, 2019).

As of now, machines lack cognitive skills (Searle, 1980; Searle, 1984; Koops et al., 2010; Gutman et al., 2012), and, it is unlikely that, at least in the near future, they will be capable of properly understanding the moral significance of their actions. Despite researchers' attempt to 'design artificial agents to act as if they [were] moral agents' and make them sensible to the 'values, ethics and legality of activities' (Allen et al., 2000; Allen et al., 2005; Lanzarone and Gobbo, 2008), a series of problems arise: the first one lies in the very definition of the ethical principles to be encoded, upon which disagreement is likely to be found; the second one is related to ambiguities connected to the use of natural language, which may lead to gaps and incongruences between what the robot is told to do, and what the designer actually intended it to do—as it is everything but trivial to translate normative statements into strings of commands; the third one is rather connected to the peculiar functioning of ethical norms, as well as many legal norms, which do not apply once and for all, but may be subject to conflicts, exceptions and balancing, which require processes of prioritization and proportionality assessments, which are far from easy to be pre-defined in a way as to be hard-coded in the machine.

Said otherwise: machines can certainly perform actions which are, in abstract terms, worthy of reactive moral attitudes; however, since they cannot engage in moral considerations, they will not qualify as moral subjects, and thus may not be attributed moral responsibility (Himma, 2009 correctly notes that all the three capacity of moral agency—rationality, ability to know the difference between right and wrong, and the ability to apply correctly these rules to certain paradigm situation that constitute the meaning of the rule –, and indeed the very concept of agency, requires the agent's consciousness).

In this sense, it is worth highlighting how the theories which accommodate artificial moral agents are often based on formal definitions and behavioristic tests that aim at proving that there is no qualitative difference between artificial and human agents. A famous example for this is the thesis offered by Floridi and Sanders, who claim that moral responsibility shall be equated to the ability to cause moral effects, which arises when an entity satisfies the formal criteria of interactivity, autonomy, and adaptability (Floridi and Sanders, 2004).

However, it has been recently demonstrated how such claims shall be read within the perspective of the machine ethics projects, and do not hold absolutely. The theoretical possibility of constructing a theory that is functional to the attribution moral agency to robots, assimilating robots and humans, does not mean that, in absolute terms, there is no significant difference between the two, nor that there is a pragmatic reason why artificial moral agency shall be constructed (Fossa, 2018).

Said theories may also be more radically challenged, for they deconstruct the notion of agency and responsibility, providing a more limited and alternative meaning to that generally accepted in the philosophical and legal discourse, yet failing to argue the reason why such an alternative proposal ought to be accepted. Said otherwise, why moral agency ought to be defined as the possibility to produce morally relevant consequences, irrespective of any identifiable intention and awareness,[10] which are instead identified as a requirement by all moral and legal paradigms, is itself to be questioned. On the one hand, their philosophical admissibility is not self-evident. On the other hand, as per their practical implications, so conceived, they are useless. Holding a machine responsible that does not fear the sanction, deprives the legal norm of its primary purpose, namely that of inducing a desired behaviour on the side of the agent.

Ultimately, RAI applications do not share human's autonomy and moral awareness necessary according to an absolute—i.e., non-instrumental or sector-specific—definition of moral agency, as the latter "cannot abstract from the very determination of ultimate ends and values, that is, of what strikes our conscience as worthy of respect and concretization" (Fossa, 2018).

RAI as *Accountable* Subjects? The Legal Notion(s) of Liability

In legal terms, being liable means to be responsible or answerable for something at law. It rests on the idea that there are specific sources of obligations, which bind one subject to do something, denoted as the object of the obligation.

In criminal matters, liability arises because of a court decision, when the prosecutor demonstrates beyond reasonable doubt that the defendant's conduct meets both the mental and the physical elements required for the offence to be punished under criminal law, and consists in fines and imprisonment, as well as other non-custodial punishments. Under western legal tradition, criminal liability has a sanctioning, as well as a re-educative aim.[11]

Civil liability rules determine who is supposed to bear the negative economic consequences arising from an accident, and under which conditions. Here, liability means "the law determining when the victim of an accident is entitled to recover losses from the injurer" (Shavell, 2007a). Typically, the party is held liable, and thence bound to compensate, that is deemed to have caused the accident, and therefore is responsible for it. Liability is established after a trial, where the claimant, who sued the wrongdoer, must prove the existence of the specific constitutive elements that ground the liability affirmed. Under English civil law, for example, to hold a person liable for negligence, the claimant needs to prove that the defendant had a duty, that she breached it, and that such breach caused an injury, resulting in recoverable damages; for instance, because the harm is not too remote a consequence of the breach (Van Gerven et al., 2000).[12]

Civil liability rules pursue three distinct functions, namely: i) ex-ante deterrence, since they aim at making the agent refrain from the harmful behavior, given that she will have to internalize the negative consequences caused; ii) ex-post compensation of the victim, as they force the person responsible for the damage to make good for the loss suffered; (iii) and ex post punishment, since the compensatory award also constitute a sanction, making sure that the infringer does not get away with the illicit behavior.

Many different theories have been elaborated to justify civil liability, as well as to shape its rules within a legal system according to specific ideologies; most of them are related to a different notion of justice. According to a retributive account of justice, the blameworthy deserve to suffer, because of the socially reprehensible character of their conduct, and liability rules shall be framed to serve as sanctions (Walen and Winter 2016). Theories of corrective justice, instead, understand tort law as a system of second-order duties (Coleman, 2003), setting obligations to make good the wrong caused by the breach of first-order duties; under this view, liability rules shall rather be elaborated and interpreted to assure that the victim is put, as much as possible, in the position she would be, had the damage not occurred. Thus, for a loss to be wrongful and worthy of being compensated, it needs to derive not from morally reprehensible conduct, but rather from a damaging violation of the victim's right.[13]

[10] The notion so conceived also denies the minimal prerequisite of *suitas*, Padovani, T (2002). *Diritto Penale*. Milano: Giuffrè. 111-112, whereby the absolute lack of any intention prevents the very possibility of assessing any (criminal) responsibility of the agent. Indeed, the latter notion builds upon and is deeply rooted in the philosophical debate in this subject matter.

[11] See art. 27 Italian Constitution: "Le pene (. . .) devono tendere alla rieducazione del condannato".

[12] For leading cases on the tort of negligence and on compensatory damages arising therefrom, see Donoghue v. Stevenson [1932] AC 532, 580; *Nettleship v Weston* [1971] 2 QB 691; *Smith v Leech Brain and Co.* [1962] 2 QB 405; *The Wagon Mound No.2* [1967] 1 AC 617 Privy Council.

[13] Under some version of this theory—developed to object other forms of liability, as developed by the school of law and economics—the principle of corrective justice that justifies the link which tort law creates between the victim and injurer, since it takes the injurer to have the duty to repair the wrongful losses that he causes, and neatly considers compensation as the primary function of liability, against that of inducing efficient behaviour.

In Law&Economics (L&E) theories, liability rules constitute economic incentives, leading agents to adopt economically efficient behaviors, which increase the overall social benefit. In this sense, paying damage is almost equal to buying the right to obtain the benefit associated with the wrong (Calabresi, 1970; Calabresi and Melamed, 1972; Shavell, 2007b; Polinsky and Shavell, 2007).

Nowadays, legal systems do not commit to only one theory of tort and justice, but rather to a combination of the three: the same normative framework will feature different models of liability rules, displaying a variety of imputation criteria (causation/remoteness, subjective elements), which in turn reflect the peculiar rationales underlying the attribution of liability.

Many tort law systems—such as the Italian one[14]—have a general rule prescribing liability for damages caused by reprehensible behaviors based on fault. This solution is moved by all the different goals defined above: not only ex-post compensation and sanction but also ex-ante deterrence, since fault-based liability incentivizes agents to adopt the standard of care necessary to avoid harmful behaviors, and thus the negative economic consequences deriving from the duty to compensate.

Sometimes, however, the defendant is held liable in tort even though she did nothing blameworthy, merely because of the particular position that the she holds towards the cause of the damage: i.e., the person who has a duty to watch over some other entity—such as the keeper, owner or user of a dangerous thing, the keeper or user of an animal—, or the person who benefits from having or using a thing, or running a specific activity.[15] The basic idea underlying the ascription of liability is that who has the economic or otherwise benefits associated with possessing or running a dangerous thing or activity, should also make sure that no damages are caused, and pay whenever this happens. This model is often associated to a strict or semi-strict liability, depending on whether the defendant may exclude his duty to compensate—i.e., by demonstrating that he took all the necessary measures to prevent the harm to occur, or by demonstrating that the latter was caused by an act of God—. The stricter the liability, the more compensation-oriented, instead of deterrence- and punishment-oriented the rationale.

Further down this line, sometimes liability is ascribed to the person who is best positioned to manage and internalize the risk, preventing its occurrence and minimizing its consequences, as well as to compensate the victim once an accident occurs. Such a model is particularly common in L&E literature (Polinsky and Shavell, 2007).

A peculiar version of this model is the so-called Risk Management Approach (henceforth RMA), which is grounded on the idea that liability should not be attributed based on considerations of fault—defined as the deviation from the desired conduct—typical of most tort law systems, but rather on the party that is best positioned to i) minimize risks and ii) acquire insurance. It moves from the basic consideration that—despite liability rules may well work as incentives or disincentives towards specific behaviors–they may not ensure sufficient and efficient incentives towards a desirable ex-ante conduct, be it a safety investment—such as in the case of producers' liability—or a diligent conduct—such as the driver's in the case of road circulation—, and that end is best attained through the adoption of the detailed ex-ante applicable regulation, such as safety regulation. According to this view, liability rules should thus be freed from the burden of incentivizing the agents towards desired conducts, and rather be shaped to ensure the maximum and most efficient compensation to the victim. In extreme cases, this could also be designed as to avoid the difficulties and burdens connected to traditional judicial adjudication, and rather be based on no-fault compensatory funds (Bertolini, 2016).

THE FUNCTIONAL PERSPECTIVE

In the previous analysis, we have clarified that for an entity to be deemed an agent, it shall be able to instantiate intentional mental states capable of directly causing performance and that for it to qualify as a moral agent, it shall display what is usually referred to as "strong autonomy," i.e., the ability to decide freely and coordinate one's action towards a chosen end, as well as the moral awareness needed for understanding the moral significance of one's actions.

In doing so, we have also explained why current RAI, conceived to complete a specific task identified by their user, shall not qualify neither as agents, absent the consciousness required for them to have intentional mental states, nor as moral agents, given that, at this stage, they have no capacity to engage in moral judgments, and lack a "strong autonomy," because they can determine how to reach the goals they are programmed to achieve, but said goals are defined by an external agent—most likely, the designer, producer or programmer—. The only moral agents involved in the functioning of the RAI application remain the humans behind it, who are responsible for its goals, its model of functioning, as well as for the very choice to grant to it a certain degree of autonomy in determining how to perform intended tasks (Putman, 1964; Bertolini, 2013; Nyholm, 2018; Dignum, 2020).

Having excluded any ontological reason why robots shall be deemed autonomous agents, thus moral and legal subjects, they shall be qualified as products: "artefacts crafted by human design and labor, for the purpose of serving identifiable human needs" (Bryson and Kime, 2011; Bertolini, 2013; Fossa, 2018). Therefore, should a robot cause any damage, ordinary product liability rules would apply. Since the latter rests on the idea the producer shall be responsible because, and as long as, he is in full control of the features and actions of the products (Bertolini, 2013), the proclaimed "responsibility gap" (Matthias, 2004; Koops et al., 2010; Calo, 2015) is then only apparent. Contract law typically allows a full-fledged autonomous and conscious human being to act—when so legitimized either by the law or by the free choice of

[14] Art. 2043 Italian Civil Code: «*Risarcimento per fatto illecito*. Qualunque fatto doloso o colposo, che cagiona ad altri un danno ingiusto, obbliga colui che ha commesso il fatto a risarcire il danno».

[15] See, e.g., Wagner, G (2015). "Comparative Tort Law," in *Comparative Tort Law*, eds. M. Reimann and R. Zimmermann (Oxford: Oxford University Press).

the party—in the name and interest of another human being, immediately modifying his legal sphere (e.g., agency). Similarly, tort law allows one party to be called in to compensate the damage caused by another subject under his supervision that only at times displays limited capacity and awareness (e.g., underage child), and in other cases is instead as autonomous as the very party obliged to pay damages (e.g., an employee). In both cases, the legal system copes with a much higher degree of autonomy—that displayed by the autonomous human agents—by imputing the legal and economic consequences of their actions to another, entirely different human being, who has a very limited—much more limited than that possessed over a machine of any sort—control over their actions.

Regardless of the complexity of its functioning, as far as the RAI application performs the tasks it was designed for, it is still under the control of the producer or the programmer: even in the case of machine-learning technologies—such as neural-based systems and genetic algorithms—the unpredictability of the learning behavior does not create any actual lack of control, but rather requires the training and the associated evolution to be included in the development phase so that the product reaches the market only when it is supposed to have learnt or perfected the skill to function safely. Should such threshold be impossible to reach, so that the machine seems not to be able to develop in a predictable way, the moral and legal responsibility for the damage caused still lies on the producer/programmer, who has a duty not to put unsafe products into the market.

What has been said so far against the alleged responsibility gap served to prove that there are no compulsory ontological reasons why ordinary product liability rules shall not apply to advanced RAI. However, it could still be the case that changes to the extant paradigm shall be made, to address the regulation of new technologies, in a way that fosters technological innovation while being respectful of and driven by the respect of European values and principles (European Commission, 2018). Social and policy considerations, as well as constitutional law, may suggest the adoption of different liability models, favoring the development of applications that are particularly valuable for society, such as prostheses or devices intended to help the otherwise disabled in their everyday tasks (Bertolini, 2015).

Likewise, current liability rules may be rethought or reformed, to better pursue the goals that they are meant to achieve (Koops et al., 2010; Bertolini, 2013; Lior, 2019; Kiršienė et al., 2021). Indeed, the Product Liability Directive—which constitutes the European framework on the issue—has recently been evaluated to assess whether it is still adequate for regulating contemporary advanced technological products. Some critical elements have been identified (Expert Group on Liability and New Technologies, 2019; Bertolini and Episcopo, 2021): the uncertainty as per the qualification of software as a product, the undesired implications deriving from the development risk defense, the cost and difficulty of exactly ascertaining the existence of a defect—in particular in design –, as well as of a causal nexus between the fact and the damage. The latter burdens the claimant substantially, discouraging litigation. Also, when advanced robotics is considered, tight human-machine interaction causes different bodies of law to overlap. Indeed, if a single task is handled together by the human agent and by a machine, when an accident occurs it might be due to the fault of the former or a defect (or malfunctioning) of the latter. Apportioning liability among the two—human agent or manufacturer—might therefore require complex factual ascertainment and articulate legal analysis. For this purpose, different approaches—such as the abovementioned Risk Management Approach—have been elaborated, which suggest modifying current product liability rules to better address the new challenges brought about by technological innovation (Bertolini, 2013; Bertolini, 2014).

The Benefits of a Functionally Attributed Electronic Personhood

Even if RAI cannot qualify as autonomous beings and, thus, there is no ontological reason why they should be considered as subjects at law, it does not mean that they may not qualify as such, because of the discretional choice of the legislator, so long as the latter is well grounded on sound policy analysis.

Indeed, the constitutive independence among the notion of personhood, agency, and responsibility in the moral and legal domains is such that functional reasons could very well justify a dissociation between the different states. For example, ad hoc legal personhood could be awarded to robots, exactly as it is granted to corporations. However, to justify this choice specific end needs to be identified, and a comparative judgment on the pros and cons of this alternative, as well as other tools, shall be considered. For example, it may be useful to attribute it to robotic applications, such as software agents to be used on capital markets which would then be registered, as to identify the limits of its allowed tasks and functions, and eventually the (physical or legal) person it is representing.

With respect to liability issues, the recognition of legal personhood would mainly serve as a liability capping method; yet it would neither necessarily change the person bearing the costs of its functioning nor the cases when compensation is awarded. Unless the robot could earn revenues from its operation, its capital would have to be provided by a human, or a corporation, standing behind it, thus not necessarily shifting the burden from the party that would bear it pursuant to existing product liability rules. Such a result could also be achieved through insurance mechanisms or with a simple damages cap. Should the robot be allowed to earn a fee for its performance, this would only constitute a tax on the user, producing an overall risk-spreading effect which could be effectively achieved otherwise, for instance through the adoption of a no-fault scheme funded by the product's users in various fashions (Bertolini, 2013; Expert Group on Liability and New Technologies, 2019). Which of the different alternatives is preferable is still a matter of correctly specifying particular circumstances, among which are the size of the market for the given application and the existence of evident failures which could be designed around through ad hoc regulation; much less would depend on the machine being weakly autonomous or even able to learn.

In this sense, we do not share the radical exclusion of legal personhood which, for instance, Bryson and colleagues make, based on the asserted undesirability of the consequences associated with such solution (Bryson et al., 2017). Indeed, the authors claim that the recognition of a legal personality—although technically possible—would be "morally unnecessary and legally troublesome": in their view, legal personhood may have some emotional or economic appeal, but difficulties in holding "electronic persons" accountable when they violate the rights of others outweigh the highly precarious moral interests that RAI's legal personhood might protect. On the contrary, we argue that, although that may be the case in some circumstances—so that the humans behind them should be held responsible under the above-described risk-management approach—, other cases may well justify the recognition of such legal status, provided that said legal personality is narrowly and functionally defined (against a one-size-fits-all approach; see, e.g., Dahiyat, 2021).

CONCLUSION

The major issue faced when discussing the possibility to attribute i) subjectivity, ii) agency, and iii) responsibility to RAI is the lack of clarity in identifying the nature of the argument that may be either ontological or functional. The two paradigms lead, in fact, to divergent considerations, and should thence always be kept profoundly distinct.

On the contrary, conclusions reached in the current debate often appear ambiguous because they tend to mix the two separate perspectives. Such lack of clarity is further advanced by the constant—and otherwise beneficial—exchange between lawyers and philosophers, who utilize those apparently similar notions with very different purposes and conceptual frameworks of reference.

While in some philosophical domains the lack of intentionality might appear insufficient an argument to exclude agency, and thence responsibility, such (de- and re-)constructions may not be transposed in the legal domain. There, intentionality serves an unavoidable purpose, that of ensuring the possibility of deterrence through regulation. The norm, by threat of a sanction, induces the desired behaviour only in those entities that are aware of their own existence, possess individual preferences, and are capable of freely coordinating their actions to achieve them. The lack of any of such elements excludes the very utility of attributing responsibility and eventually sanctioning the transgressor.

At the same time, the legal system is well structured to deal with the need to impose legal effects produced by the behaviour of a subject onto another one, despite the former being fully capable of determining himself autonomously, so long as the latter identifies the ends to be achieved.

If we look at the specificities of the legal system, it is then objectively observable that there are no ontological grounds to determine the attribution of subjectivity, rights, duties, and obligations to machines. Nothing in the way the machine is designed, functions, or is justifies the legitimacy of such attributions, rather excludes them.

However, a functional analysis may lead to different conclusions so long as adequate purely legal arguments could be identified. That may only be achieved with respect to i) specific classes of applications, ii) when a technical—in a legal perspective—need is identified, that is best pursued through the attribution of rights and obligations to the machine itself, rather than the humans behind it. In such a sense, the attribution of agency, subjectivity or responsibility would not follow the acknowledgment of a special nature of the RAI, but of a legal need for—separately or jointly—a) simplification of legal relations, b) traceability, registration and transparency of the entity and of those possessing interests in it and in its operation, c) segregation of assets and limitations of responsibility.

Based on those considerations, conclusions such as those reached by the EU parliament in its recent proposal on the regulation of civil liability for AI—whereby since "[. . .] all physical or virtual activities, devices or processes that are driven by AI systems [. . .] are nearly always the result of someone building, deploying or interfering with the systems [. . .] it is not necessary to give legal personality to AI-systems" (European Parliament, 2020 introduction n° 7)—appear excessively broad and thence unjustified. Said otherwise, the mere circumstance that all legal relations revolving around corporations could be described and regulated through bundles of contracts, does not justify *per se* the exclusion of the utility of legal personhood. The reason such a conclusion is flawed in a technical legal perspective has to do with the technology-neutral approach that proposal attempts to maintain, presenting a uniform regulation for applications and use cases that are extremely different one from the other, and that today would be addressed by entirely different branches of the legal ordering (such as capital markets, traffic accidents, medical or professional malpractice, to name a few).

It is indeed certain that the cases in favor of the direct attribution of liability to a RAI application need to be individually justified, yet that debate belongs entirely to the technical-legal domain and has no bearing nor implications on the acknowledgment of the machine as an entity deserving moral standing and the attribution of individual rights.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AB is primarily responsible for §§3, 4.4, 5, 5.1; FE is primarily responsible for §§ 2, 4.1, 4.2, 4.3. Both authors equally contributed to §§1, 2, and 6.

FUNDING

The authors declare that the research was conducted under the Jean Monnet Project EURA "European Centre of excellence on the Regulation of Robotics and AI".

REFERENCES

Alexy (1978). *Theorie der juristischen Argumentation. Die Theorie des rationalen Diskurses als Theorie der juristischen Begründung (trad. it. Teoria dell'argomentazione giuridica. La teoria del discorso razionale come motivazione giuridica*. Milano, Giuffrè: Frankfurt a.M.

Allen, C., Smit, I., and Wallach, W. (2005). Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. *Ethics Inf. Technol.* 7, 149–155. doi:10.1007/s10676-006-0004-4

Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *J. Exp. Theor. Artif. Intelligence* 12, 251–261. doi:10.1080/09528130050111428

Anderson, M., and Anderson, S. (2011). *Machine Ethics*. Cambridge, United Kingdom: Cambridge Univ. Press.

Andreotta, A. J. (2021). The Hard Problem of AI Rights. *AI & Soc* 36, 19–32. doi:10.1007/s00146-020-00997-x

Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.

Aristotle (1985). *The Nicomachean Ethics*. Indianapolis: Hackett Publishing.

Austin, J., Hart, H. L. A., and Austin, J. (1954). *The Province of Jurisprudence Determined and, the Uses of the Study of Jurisprudence*. New York: New York: Noonday Press.

Balkin, J. M. (2015a). The Path of Robotic Law. *Calif. L. Rev. Circuit* 6, 45–60.

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-Temporality in Action. *Adaptive Behav.* 17, 367–386. doi:10.1177/1059712309343819

Basl, J., Bowen, J., Ai, T. O. H. O. E. O., Markus, D., Dubber, F. P., and Sunit, D. (2020). in *The Oxford Handbook of Ethics of AI,* eds. M. D. Dubber, F. P. Pasquale, and S. Das.

Basl, J. (2014). Machines as Moral Patients We Shouldn't Care about (Yet): The Interests and Welfare of Current Machines. *Philos. Technol.* 27, 79–96. doi:10.1007/s13347-013-0122-y

Bennett, B., and Daly, A. (2020). Recognising Rights for Robots: Can We? Will We? Should We? *L. Innovation Tech.* 12, 60–80. doi:10.1080/17579961.2020.1727063

Bentham, J. (1823). *An Introduction to the Principles of Morals and Legislation*. London: E. Wilson.

Bertolini, A., and Episcopo, F. (2021). The Expert Group's Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies: a Critical Assessment. *Eur. J. Risk Regul.* 12, 644–659. doi:10.1017/err.2021.30

Bertolini, A. (2016). Insurance and Risk Management for Robotic Devices: Identifying the Problems. *Glob. Jurist* 16, 291–314. doi:10.1515/gj-2015-0021

Bertolini, A. (2015). Robotic Prostheses as Products Enhancing the Rights of People with Disabilities. Reconsidering the Structure of Liability Rules. *Int. Rev. L. Comput. Tech.* 29, 116–136. doi:10.1080/13600869.2015.1055659

Bertolini, A. (2014). "Robots and Liability - Justifying a Change in Perspective," in *Rethinking Responsibility in Science and Technology*. Editors F. Battaglia, J. Nida-Rümelin, and N. Mukerji (Pisa: Pisa University Press), 143–166.

Bertolini, A. (2013). Robots as Products: The Case for a Realistic Analysis of Robotic Applications and Liability Rules. *L. Innovation Tech.* 5, 214–247. doi:10.5235/17579961.5.2.214

Bobbio, N. (1969). "Sanzione," in *Novissimo Digesto* (Torino: UTET).

Boden, M. A. (1996). *The Philosophy of Artificial Life*. Oxford: Oxford University Press.

Brożek, B., and Janik, B. (2019). Can Artificial Intelligences Be Moral Agents? *New Ideas Psychol.* 54, 101–106.

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the People: the Legal Lacuna of Synthetic Persons. *Artif. Intell. L.* 25, 273–291. doi:10.1007/s10506-017-9214-9

Bryson, J. J., and Kime, P. P. (2011). "Just an Artifact: Why Machines Are Perceived as Moral Agents," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence: Barcelona, Catalonia, Spain, 16–22 July 2011*. Editor T. Walsh (Menlo Park, CA, USA: AAAI Press), 1641–1646.

Bunge, M. (1977). Towards a Technoethics. *The Monist* 60, 96–107. doi:10.5840/monist197760134

Calabresi, G., and Melamed, A. D. (1972). Property Rules, Liability Rules, and Inalienability: One View of the Cathedral. *Harv. L. Rev.* 85, 1089. doi:10.2307/1340059

Calabresi, G. (1970). *The Cost of Accidents*. New Haven: Yale University Press.

Calo, R. (2015). Robotics and the Lessons of Cyberlaw. *Calif. L. Rev.* 103, 513–563.

Chrisley, R. (2008). Philosophical Foundations of Artificial Consciousness. *Artif. Intelligence Med.* 44, 119–137. doi:10.1016/j.artmed.2008.07.011

Christman, J. (2018). *Autonomy in Moral and Political Philosophy*.

Coeckelbergh, M. (2010). Robot Rights? towards a Social-Relational Justification of Moral Consideration. *Ethics Inf. Technol.* 12, 209–221. doi:10.1007/s10676-010-9235-5

Coleman, J. L. (2003). *The Practice of Principle : In Defence of a Pragmatist Approach to Legal Theory. Oxford*. Oxford: Oxford University Press.

Dahiyat, E. A. R. (2021). Law and Software Agents: Are They "Agents" by the Way? *Artif. Intell. L.* 29, 59–86. doi:10.1007/s10506-020-09265-1

Danaher, J. (2020). Robot Betrayal: a Guide to the Ethics of Robotic Deception. *Ethics Inf. Technol.* 22, 117–128. doi:10.1007/s10676-019-09520-3

Davidson, D. (1963). Actions, Reasons, and Causes. *J. Philos.* 60, 685–700. doi:10.2307/2023177

De Jong, R. (2020). The Retribution-gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm. *Sci. Eng. Ethics* 26, 727–735. doi:10.1007/s11948-019-00120-4

De Pagter, J. (2021). Speculating about Robot Moral Standing: On the Constitution of Social Robots as Objects of Governance. *Front. Robot AI* 8, 769349. doi:10.3389/frobt.2021.769349

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little Brown & Co.

Dennett, D. C. (2018). Facing up to the Hard Question of Consciousness. *Phil. Trans. R. Soc. B* 373, 20170342. doi:10.1098/rstb.2017.0342

Dickson, J. (2021). *Ours Is a Broad Church: Indirectly Evaluative Legal Philosophy as a Facet of Jurisprudential Inquiry*. Taylor & Francis, 207–230.

Dignum, V. (2020). "Responsibility and Artificial Intelligence," in *The Oxford Handbook of Ethics of AI*. Editors M. D. Dubber, F. Pasquale, and S. Das (Oxford: Oxford University Press), 215–231. doi:10.1093/oxfordhb/9780190067397.013.12

Dworkin, R. (1986). *Law's empire*. Cambridge, Mass: Belknap Press.

Dworkin, R. (1977). *Taking Rights Seriously*. Cambridge Mass: Harvard Univeristy Press.

Endicott, T. (2000). *Vagueness in Law*. Oxford: Oxford University Press.

Eshleman, A. (2016). "Moral Responsibility," in *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)* Editor. E. N. Zalta. Available at https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/.

European Commission (20182018). "Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe," in *COM* (Brussels: European Commission), 237.

European Parliament (2020). *Civil Liability Regime for Artificial Intelligence. European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence (2020/2014(INL)*. Brussels: European Parliament.

European Parliament (2017). *European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics*. European Parliament. 2015/2103(INL).

Expert Group on Liability and New Technologies (2019). *Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies*. Brussels: European Commission.

Falzea, A. (1960). "Voce «Capacità (Teoria gen.)," in *Enciclopedia del diritto* (Milano: Giuffrè).

Finnis, J. (2011). *Natural Law and Natural Rights*. Oxford University Press.

Finnis, J. (2020). "Natural Law Theories," in *The Stanford Encyclopedia of Philosophy*. Editor E. Zalta.

Floridi, L. (2014). "Artificial Agents and Their Moral Nature," in *The Moral Status of Technical Artefacts*. Editors P. Kroes and P.-P. Verbeek (Dordrecht: Springer Netherlands), 185–212. doi:10.1007/978-94-007-7914-3_11

Floridi, L., and Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14, 349–379. doi:10.1023/b:mind.0000035461.63578.9d

Fossa, F. (2021). Artificial agency and the Game of Semantic Extension. *Interdiscip. Sci. Rev.* 46, 440–457. doi:10.1080/03080188.2020.1868684

Fossa, F. (2018). Artificial Moral Agents: Moral Mentors or Sensible Tools? *Ethics Inf. Technol.* 20, 115–126. doi:10.1007/s10676-018-9451-y

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *J. Philos.* 68, 5–20. doi:10.2307/2024717

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *J. Conscious. Stud.* 23, 11–39.

Gabriel, M. (2021). "Could a Robot Be Conscious? Some Lessons from Philosophy," in *Robotics, AI, and Humanity: Science, Ethics, and Policy*. Editors J. Von Braun, M. S. Archer, G. M. Reichberg, and M. Sánchez Sorondo (Cham: Springer International Publishing), 57–68. doi:10.1007/978-3-030-54173-6_5

Gellers, J. C. (2021). *Rights for Robots. Artificial Intelligence, Animal and Environmental Law*. London: Routledge.

Gert, B., and Gert, J. (2020). "The Definition of Morality," in *The Stanford Encyclopedia of Philosophy*. Editor E. N. Zalta.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Gogoshin, D. L. (2021). Robot Responsibility and Moral Community. *Front. Robot AI* 8, 768092. doi:10.3389/frobt.2021.768092

Gordon, J.-S. (2021). Artificial Moral and Legal Personhood. *AI Soc.* 36, 457–471. doi:10.1007/s00146-020-01063-2

Green, L., and Adams, T. A. (2019). "Legal Positivism," in *The Stanford Encyclopedia of Philosophy*. Editor E. Zalta.

Gunkel, D. J. (2020). Mind the gap: Responsible Robotics and the Problem of Responsibility. *Ethics Inf. Technol.* 22, 307–320. doi:10.1007/s10676-017-9428-2

Gunkel, D. J. (2018). *Robot Rights*. mit Press.

Gunkel, D. J. (2012). *The Machine Question*.

Gunkel, D. J., and Wales, J. J. (2021). Debate: what Is Personhood in the Age of AI? *AI Soc.* 36, 473–486. doi:10.1007/s00146-020-01129-1

Gutman, M., Rathgeber, B., and Syed, T. (2012). "Action and Autonomy: A Hidden Dilemma in Artificial Autonomous Systems," in *Robo- and Informationethics. Some Fundamentals* (LIT Verlag Münster: M. Decker & M. Gutman.Zürich: Lit), 231–256.

Hart, H. L. A. (2012). *The Concept of Law*. Oxford: Oxford University Press.

Haselager, W. F. G. (2005). Robotics, Philosophy and the Problems of Autonomy. *P&C* 13, 515–532. doi:10.1075/pc.13.3.07has

Hershovitz, S. (2015). The End of Jurisprudence. *Yale L. J.* 124, 1160–1205.

Himma, K. E. (2009). Artificial agency, Consciousness, and the Criteria for Moral agency: what Properties Must an Artificial Agent Have to Be a Moral Agent? *Ethics Inf. Technol.* 11, 19–29. doi:10.1007/s10676-008-9167-5

Howard, D., and Muntean, I. (2017). "Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency," in *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*. Editor T. M. Powers (Cham: Springer International Publishing), 121–159. doi:10.1007/978-3-319-61043-6_7

Iannì, A., and Monterossi, M. W. (2017). Artificial Autonomous Agents and the Question of Electronic Personhood: a Path between Subjectivity and Liability. *Griffith L. Rev.* 26, 563–592. doi:10.1080/10383441.2017.1558611

Joshua, C. G. (2021). *Rights for Robots : Artificial Intelligence, Animal and Environmental Law*. London: Routledge.

Jost, T. S. (2002). Rights of Embryo and Foetus in Private Law. *Am. J. Comp. L.* 50. doi:10.2307/841064

Jowitt, J. (2021). Assessing Contemporary Legislative Proposals for Their Compatibility with a Natural Law Case for AI Legal Personhood. *AI Soc.* 36, 499–508. doi:10.1007/s00146-020-00979-z

Kelsen, H. (2005). *Pure Theory of Law*. Clark, N.JLawbook Exchange.

Kingwell, M. (2020). "The Oxford Handbook of Ethics of AI," in *The Oxford Handbook of Ethics of AI*. Editors M. D. Dubber, F. Pasquale, and S. Das (Oxford: Oxford University Press), 326–342.

Kiršienė, J., Gruodytė, E., and Amilevičius, D. (2021). From Computerised Thing to Digital Being: mission (Im)possible? *AI SOCIETY* 36, 547–560.

Koops, B.-J., Hildebrandt, M., and Jaquet-Chiffelle, D.-O. (2010). Bridging the Accountability Gap: Rights for New Entities in the Information Society? *Minn. J. L. Sci. Technol.* 11, 497–561.

Kraakman, R., Armour, J., Davies, P., Enriques, L., Hansmann, H., Hertig, G., et al. (2017). *The Anatomy of Corporate Law: A Comparative and Functional Approach*. Oxford: Oxford University Press.

Kurki, V. a. J. (2019). *A Theory of Legal Personhood*. Oxford, United Kingdom: Oxford University Press.

Lanzarone, G. A., and Gobbo, F. (2008). "Is Computer Ethics Computable?," in *Conference Proceedings of ETHICOMP 2008: Living, Working and Learning beyond Technology*. Editor T. W. E. A. Bynum (Mantova: Tipografia Commerciale), 530.

Lior, A. (2019). AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy. *Mitchell Hamline L. Rev.* 46, 1043–1102.

Loh, J. (2019). Responsibility and Robot Ethics: A Critical Overview. *Philosophies* 4, 58. doi:10.3390/philosophies4040058

Marmor, A., and Sarch, A. S. (2019). "The Nature of Law," in *The Stanford Encyclopedia of Philosophy*. Editor E. Zalta.

Martínez, E., and Winter, C. (2021). Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection. *Front. Robotics AI* 8. doi:10.3389/frobt.2021.788355

Matthias, A. (2010). *Automaten als Träger von Rechten*. Berlin: Logos Verlag.

Matthias, A. (2008). "From Coder to Creator. Responsibility Issues in Intelligent Artifact Design," in *Handbook of Research in Technoethics*. Editors R. Luppicini and R. A. Hersher.

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics Inf. Technol.* 6, 175–183. doi:10.1007/s10676-004-3422-1

Mcnally, P., and Inayatullah, S. (1988). The Rights of Robots: Technology, Culture and Law in the 21st Century. *Futures* 20. doi:10.1016/0016-3287(88)90019-5

Nagel, T. (1974). What Is it like to Be a Bat? *Phil. Rev.* 83, 435–450. doi:10.2307/2183914

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci. *Sci. Eng. Ethics* 24, 1201–1219. doi:10.1007/s11948-017-9943-x

Nyholm, S. (2020). *Humans and Robots: Ethics, agency, and Anthropomorphism*. Lanham, MD: Rowman & Littlefield Publishers.

Osborne, D. S. (2020). Personhood for Synthetic Beings: Legal Parameters and Consequences of the Dawn of Humanlike Artificial Intelligence. *Santa Clara High Tech. L. J.* 37, 257–300.

Padovani, T. (2002). *Diritto Penale*. Milano: Giuffrè.

Pagallo, U. (2018b). Vital, Sophia, and Co.-The Quest for the Legal Personhood of Robots. *Information* 9, 230. doi:10.3390/info9090230

Pagallo, U. (2018a). Vital, Sophia, and Co.—The Quest for the Legal Personhood of Robots. *Information* 9.

Palmerini, E., and Bertolini, A. (2016). "Liability and Risk Management in Robotics," in *Digital Revolution: Challenges for Contract Law in Practice*. Editors R. Schulze and D. Staudenmayer (Baden-Baden: Nomos), 225–260. doi:10.5771/9783845273488-225

Polinsky, M. A., and Shavell, S. (2007). *Handbook of Law and Economics*. North-Holland.

Powell, D. (2020). Autonomous Systems as Legal Agents: Directly by the Recognition of Personhood or Indirectly by the Alchemy of Algorithmic Entities. *Duke L. Tech. Rev.* 18, 306–331.

Prescott, T. J. (2017). Robots Are Not Just Tools. *Connect. Sci.* 29, 142–149. doi:10.1080/09540091.2017.1279125

Purves, D., Jenkins, R., and Strawser, B. J. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethic Theor. Moral Prac* 18, 851–872. doi:10.1007/s10677-015-9563-y

Putman, H., and Putnam, H. (1964). Robots: Machines or Artificially Created Life? *J. Philos.* 61, 668–691. doi:10.2307/2023045

Raz, J. (2009). *The Authority of Law : Essays on Law and Morality*. New York: Oxford.

Rest, J. R. (1986). *Moral Development: Advances in Research and Theory*. New York: Praeger Publishers.

Richards, N. M., and Smart, W. D. (2016). "How Should the Law Think about Robots?," in *Robot Law* (Cheltenham, United Kingdom: Edward Elgar Publishing).

Sacco, R. (1991). Legal Formants: A Dynamic Approach to Comparative Law (Installment II of II). *Am. J. Comp. L.* 39, 343–401. doi:10.2307/840784

Santoni De Sio, F., and Van Den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Front. Robot AI* 5, 15. doi:10.3389/frobt.2018.00015

Schlosser, M. (2015). "Agency," in *The Stanford Encyclopedia of Philosophy (Fall 2015 Edition)*. Editors E. N. Zalta, U. Nodelman, C. Allen, and R. L. Anderson (Stanford, CA: Stanford University).

Schröder, W. M. (2021). "Robots and Rights: Reviewing Recent Positions in Legal Philosophy and Ethics," in *Robotics, AI, and Humanity: Science, Ethics, and*

Policy. Editors J. Von Braun, M. S. Archer, G. M. Reichberg, and M. Sánchez Sorondo (Cham: Springer International Publishing), 191–203. doi:10.1007/978-3-030-54173-6_16

Schwitzgebel, E., and Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Stud. Philos.* 39, 98–119. doi:10.1111/misp.12032

Searle, J. R. (1980). Minds, Brains, and Programs. *Behav. Brain Sci.* 3, 417–424. doi:10.1017/s0140525x00005756

Searle, J. R. (1984). *Minds, Brains, and Science*. Cambridge, Mass: Harvard University Press.

Serafimova, S. (2020). Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement. *Humanit Soc. Sci. Commun.* 7, 119. doi:10.1057/s41599-020-00614-8

Seymour, J. (2002). The Legal Status of the Fetus: an International Review. *J. L. Med* 10, 28–40. doi:10.1080/0907676x.2002.9961430

Shavell, S. (2007b). "Chapter 2 Liability for Accidents," in *Handbook of Law and Economics*. Editors A. M. Polinsky and S. Shavell (Amsterdam: North Holland - Elsevier), 139–182. doi:10.1016/s1574-0730(07)01002-x

Shavell, S. (2007a). "Liability for Accidents," in *Handbook of Law and Economics*. Editors A. M. Polinsky and S. Shavell (Amsterdam: Elsevier), 142.

Singer, W. (2021). "Differences between Natural and Artificial Cognitive Systems," in *Robotics, AI, and Humanity: Science, Ethics, and Policy*. Editors J. Von Braun, M. S. Archer, G. M. Reichberg, and M. Sánchez Sorondo (Cham: Springer International Publishing), 17–27. doi:10.1007/978-3-030-54173-6_2

Sirena, P. (2020). Introduction to Private Law. *Il Mulino*.

Solum, L. B. (1992). Legal Personhood for Artificial Intelligences. *N.C. L. Rev.*, 1231. Available at http://scholarship.law.unc.edu/nclr/vol70/iss4/4.

Stahl, B. C. (2006). Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or agency. *Ethics Inf. Technol.* 8, 205–213. doi:10.1007/s10676-006-9112-4

Stanzione, P. (1988). "Voce «Capacità I) Diritto Privato," in *Enciclopedia Giuridica* (Bologna-Roma: Zanichelli-Foro it).

Strawson, P. F. (1962). "Freedom and Resentment," in *Proceedings of the British Academy*. Editor G. Watson (Oxford: Oxford University Press), 1–25.

Taylor, C. (1977). "What Is Human Agency?," in *The Self: Psychological and Philosophical Issues*. Editor T. Michel (Oxford: Blackwell), 103–135.

Teubner, G. (2006). Rights of Non-humans? Electronic Agents and Animals as New Actors in Politics and Law. *J. L. Soc.* 33, 497–521. doi:10.1111/j.1467-6478.2006.00368.x

Thiel, U. (2011). *The Early Modern Subject : Self-Consciousness and Personal Identity from Descartes to Hume*. New York: Oxford University Press.

Toulmin, S. (1964). *The Uses of Arguments*. Cambridge: Cambridge University Press.

Turing, A. M. (1950). I.-Computing Machinery and Intelligence. *Mind* LIX, 433–460. doi:10.1093/mind/lix.236.433

Turner, J. (2019). *Robot Rules: Regulting Artiicial Intelligence*. Berlin: Springer.

Van Gerven, W., Lever, J., and Larouche, P. (2000). *Tort Law*. Oxford: Hart Publishing.

Wagner, G. (2015). "Comparative Tort Law," in *Comparative Tort Law*. Editors M. Reimann and R. Zimmermann (Oxford: Oxford University Press).

Wagner, G. (2019). Robot, Inc.: Personhood for Autonomous Systems? *Fordham L. Rev.* 88

Waldron, J. (1994). Vagueness in Law and Language: Some Philosophical Issues. *Calif. L. Rev.* 82, 509. doi:10.2307/3480971

Walen, A. (2016). "Retributive Justice," in *The Stanford Encyclopedia of Philosophy*. Editor E. Zalta. (URL = Available at: https://plato.stanford.edu/archives/win2016/entries/justice-retributive/>.

Wallach, W., and Allen, C. (2009a). *Teaching Robots Right from Wrong*. Oxford: Oxford University Press.Moral Machines

Wallach, W., and Allen, C. (2009b). *Moral Macines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Wheeler, M. (2020). "Autonomy," in *The Oxford Handbook of Ethics of AI*. Editors M. D. Dubber, F. Pasquale, and S. Das (Oxford: Oxford University Press), 333–358. doi:10.1093/oxfordhb/9780190067397.013.22

Wren, T. E., Edelstein, W., and Nunner-Winkler, G. (1990). *The Moral Domain: Essays in the Ongoing Discussion Betweeen Philosophy and the Social Sciences*. Mit Press.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership