Computer Ethics - Philosophical Enquiry (CEPE) Proceedings

Volume 2019 *CEPE 2019: Risk & Cybersecurity* Article 13

5-29-2019

Human Supremacy as Posthuman Risk

Daniel Estrada
New Jersey Institute of Technology

Human supremacy as posthuman risk

Daniel Estrada
New Jersey Institute of Technology

Abstract

Human supremacy is the widely held view that human interests ought to be privileged over other interests as a matter of public policy. Posthumanism is an historical and cultural situation characterized by a critical reevaluation of anthropocentrist theory and practice. This paper draws on Rosi Braidotti's critical posthumanism and the critique of ideal theory in Charles Mills and Serene Khader to address the use of human supremacist rhetoric in AI ethics and policy discussions, particularly in the work of Joanna Bryson. This analysis leads to identifying a set of risks posed by human supremacist policy in a posthuman context, specifically involving the classification of agents by type.

Keywords: *AI Ethics, Posthumanism, Anthropocentrism, Ideal Theory, Critical Robot Studies*

Against the backdrop of numerous political scandals, ethical violations, and calls for regulatory oversight in the field of artificial intelligence (Whittaker et al., 2018), the rhetorical framework of the "human" has become an increasingly visible shorthand for industry and public policy projects to signal a concern for safety, ethical integrity, and the responsible use of the AI. Several recent public policy proposals on AI bear titles such as "AI for Humanity" in France[1]; "HumaneAI" in the EU[2]; "AI4People", Luciano Floridi's recent proposal describing "An Ethical Framework for a Good AI Society" (Floridi et al., 2018), and Stanford's "Institute for Human-Centered AI,"[3] whose welcome page proudly proposes that "AI is to serve the collective needs of humanity."

In these contexts, centering the "human" as the explicit focus of normative concern evokes the image of an inclusive framework of shared values and common interests to guide the collective use of AI. More subtly, these proposals consistently frame AI in the role of a *servant* to human interests. Unsurprisingly, these proposals don't define the scope or content of the category "human," how membership in this category is to be determined, or what underpins its role as a focal norm in AI policy. And yet given that AI is regularly used to target, manipulate, incarcerate, and exploit vulnerable populations (witness the use of facial recognition technologies in policing, military, and security; election tampering and voter manipulation on social media; software that automates criminal sentencing, loan approval, hiring decisions, and so on; see (Angwin, Larson, Mattu, & Kirchner, 2016; Asaro, 2016; Benjamin, 2016;

[1] https://www.aiforhumanity.fr
[2] https://www.humane-ai.eu/
[3] https://hai.stanford.edu/

Buolamwini & Gebru, 2018; Eubanks, 2018; Noble, 2018; Spiel, Keyes, & Barlas, 2019; Williams, 2019b) and many others), how can we trust that policies which center the "human" will also center *us*? It is difficult to see how technologies used to expand, reinforce, and make more profitable these pervasive institutions of violence and oppression could operate against a shared background of values common to all humanity. Rosi Braidotti quotes Tony Davies: "All humanisms, until now, have been imperial" (Braidotti, 2013, p. 15). From this perspective, the attempt to signal a commitment to integrity by appeal to the category "human" reads as more of the empty ethics washing that has come to characterize the field (Metzinger, 2019).

Absent from these human-centered proposals is any engagement with the decades of sustained scholarship in feminist, postcolonial, and critical race theory (Deckha, 2012; D. Haraway, 1989, 1991; Hayles, 2008; Mills, 2011; E. W. Said, 2004), animal rights and environmental ethics (Belcourt, 2015; Gaard, 2011; Katz, 2000; Steiner, 2010); STS, HCI, and design theory (Kera, Block, & Link, 2009; Latour, 2003; Thomas, Remy, & Bates, 2017; Valentin, 2014), and related fields that have developed systematic critiques of anthropocentrism and the politics of the "human." Among the important insights of this diverse literature is the recognition that a superficial appeal to inclusive universalism can be used to justify and provide cover for narrowly self-serving, exclusive, or imperialist practices (Giraldo, 2016; Khader, 2018). As Charles Mills puts the point, "historically and still currently, most humans were not and are not socially recognized persons, or, more neatly and epigrammatically put: *most persons are non-persons*" (Mills, 2011). Confronting such duplicitous ideologies presents difficult conceptual, rhetorical, and practical challenges, suggesting that care should be taken in the use of universalizing language (if it is used at all). The uncritical deployment in AI policy of human-centered rhetoric as a pretense to ethical integrity speaks not only as a tacit endorsement of its imperialist undertones, but more loudly as utter disregard for those scholars and activists who have been consistently engaged with the "human" as a normative ground.

This paper seeks to correct these omissions and provoke the AI community to adopt a more reflective, informed, and critical perspective on the way human-centered rhetoric can function as a cheap proxy for ethical integrity. To these ends we engage the work of Joanna Bryson, prominent scholar, public speaker, and policy consultant in AI ethics, and author of the unfortunate paper "Robots Should Be Slaves" (Bryson, 2010). In this and several other essays (Bryson & Kime, 2011; Bryson, Diamantis, & Grant, 2017; Bryson, 2018a, 2018b), Bryson and colleagues construct a vision of the ethical use of AI that renders these technologies as an explicit social underclass modelled after the historical institutions of slavery. This rendering of the "human" as categorically dominant over artificial agents *in virtue of our kind* is the target of this analysis of "human supremacy." Reading Bryson through the wealth of critical scholarship on humanism and anthropocentrism, especially as discussed in critical race and postcolonial studies, does not merely raise a set of objections to the tone and content of her work. It also offers a case study on the relative ease with which ideologies of oppression can develop from what might seem like an innocuous commitment to an ethic and ontology that puts the "human" first.

Bryson is by no means alone in her explicit endorsement of the institutionalized slavery of machines (Petersen 2007, 2012). Oscar Wilde anticipates the position as

early as 1891: “Human slavery is wrong, insecure, and demoralising. On mechanical slavery, on the slavery of the machine, the future of the world depends” (Wilde, 1891). Ruha Benjamin presents a striking example of human supremacist rhetoric in an article from *Mechanix Illustrated* from 1965 that predicts “Slavery will be back! We’ll all have personal slaves again… Don’t be alarmed. We mean robot ‘slaves’.” Benjamin notes that “It goes without saying that readers, so casually hailed as “we,” are not the descendants of those whom Lincoln freed” (Benjamin, 2019, p. 56). This helps locate the matrix of oppression, artificial agency, and group identity that we wish to address with a critical evaluation of human supremacy in AI. Benjamin continues: “For those of us who believe in a more egalitarian notion of power, of collective empowerment without domination, how we imagine our relation to robots offers a mirror for thinking through and against race as technology” (Benjamin, 2019, p. 56ff). This paper is presented in solidarity with those communities who share this commitment to collective empowerment.

The paper is organized as follows. We begin by introducing the term “human supremacy” as it appears in the animal advocacy literature, and we engage the conceptual and interpretive challenges the term invites in its application to AI ethics. We go on to sketch Bryson’s human-centered approach to AI ethics as a paradigm case of human supremacy, addressing its theoretical grounding and consequences for policy. With Bryson’s views unpacked, we then turn to two resources to understand them: Rosi Braidotti’s discussion of reactionary posthumanism in Martha Nussbaum, and the critique of ideal theory in Charles Mills and Serene Khader. This scaffolding reveals the ideological and institutional foundations for Bryson’s position, and points to an alternative approach that emphasizes the nonideal conditions within which subjectivity and community operate. The paper closes with reflections on the risks of human supremacist politics in a posthuman age, specifically concerning the classification of agents by type.

Human Supremacy in Animal Rights and AI Ethics

Human supremacy is the view that human interests ought to be systematically privileged over other interests as a matter of public policy. The term derives from activist-scholars in animal rights and environmental ethics (Crist, 2017; Lupinacci, 2015; Steiner, 2010) who object to anthropocentric policies that neglect the welfare and integrity of nonhuman biological and ecological systems. Similar terms can be found in, for instance, Mary Midgley’s “human chauvinism” or “exclusive humanism” (Midgley, 2003). However, the term “human supremacy” draws on a deliberately provocative analogy to white supremacy, that pervasive system of racist structural power and oppression which systematically privileges the interests of people identified as “white” relative to people who, for various historical and sociopolitical reasons, are not so identified. Analogously, human supremacy names those practices which systematically privilege the “human” relative to the “nonhuman”. Such ideologies of oppression cannot be easily reduced to the prejudiced beliefs or attitudes that some individuals or groups hold towards others. The vocabulary of oppression highlights the structural, institutional,

and material realities within which some social groups are systematically attacked, exploited, and marginalized relative to others. (Frye, 1983; Young, 1988).

The problematic analogy between white supremacy and the racist oppression of humans on one hand and human supremacy and the anthropocentric oppression of animals and other nonhumans on the other has been addressed in critical race studies, critical animal studies, and ecofeminist literatures (Armstrong, 2002; Gaard, 2011; Nocella, 2012; Nocella, White, & Cudworth, 2015; Wise, 2005). These sources emphasize that the comparison between oppressed people and industrial livestock is deeply insensitive to the history of racialized chattel slavery that operated on this analogy[4]. Critical animal studies scholar Anthony Nocella (2012) argues that members of the animal advocacy movement rarely share a common experience of oppression, either as a community or with the animals they advocate for. This points to an important disanalogy between the (nonhuman) animal rights movement and the ongoing struggles against racist, sexist, ableist, and colonialist oppression. A nonhuman animal might maul its captor and escape its bonds, but that animal cannot engage directly in political resistance without animal activists working on their behalf. The resistance to the oppression of humans stands in stark contrast; Nocella quotes political prisoner and former member of the Black Panther Party Jalil Muntaqim, "We are our own liberators" (Nocella, 2012, p. 148).

To avoid the "white savior complex" the phrase "animal liberation" implies, Nocella suggests thinking instead of "animal justice," and appreciating how interconnected structures of oppression and domination reveals that "fighting for human animal rights *is* fighting for nonhuman animal rights" (Nocella, 2012, p. 150). Similarly, Gaard (2011) points to efforts in ecofeminist thought that foreground the intersections of race, gender, class, and ecology, for example in, "industrialized animal food production and its reliance on undocumented immigrant workers (who risk deportation if they report their hazardous workplace conditions)" (Gaard, 2011, p. 36). Nocella offers several recommendations to integrate the work of animal advocacy more deeply with other struggles to end oppression, including centering the work of people of color who are engaged in social justice and animal advocacy; challenging one's own whiteness, domination, and elitism[5]; and resisting the comparison between

---

[4] Reacting to a series of photo campaigns from the People for the Ethical Treatment of Animals (PETA) with titles like "Are Animals the New Slaves?" and "Holocaust on your plate", white anti-racist author Tim Wise explains this insensitivity by appeal to the white privilege of many animal rights activists:

> "That PETA can't understand what it means for a black person to be compared to an animal, given a history of having been thought of in exactly those terms, isn't the least bit shocking. After all, the movement is perhaps the whitest of all progressive or radical movements on the planet, for reasons owing to the privilege one must possess in order to focus on animal rights as opposed to, say, surviving oneself from institutional oppression." (Wise, 2005)

[5] While I do not identify as white, I am a straight cisgendered able-bodied man with an education and a full-time teaching position at a public technical institute, and these advantages put me in a position of privilege relative to many oppressed and marginalized people. These advantages have allowed me the opportunity to address the social status of artificial agents as a philosophical and scholar-activist project, an opportunity made possible by the very same social structures that are systematically targeting Latinx members of my communities in Southern California for detention and deportation. I'd like to acknowledge that this research was done during the tragic

forms of oppression "if that comparison is not met with action, and is not examined for the purpose of understanding the oppressor" (Nocella, 2012, p. 152).

Nocella's advice applies equally to the potentially insensitive comparisons the term "human supremacy" invites in the context of artificial agents and AI[6]. It is critical to bear in mind that the that the discourse around the "human" arises in the AI literature at the same time as egregious ethical failures in both industry and public policy that disproportionately impact the lives of people who have already been marginalized and exploited by racism and white supremacy, sexism and patriarchy, transphobia, ableism, nationalist xenophobia, and other forms of systemic oppression (Bardzell & Bardzell, 2015; Irani, Vertesi, Dourish, Philip, & Grinter, 2010; Keyes, Hoy, & Drouhard, 2019). A critical inquiry into human supremacy in AI and the systematic oppression of robots, or what might be called *critical robot studies*, is motivated by an attempt to expose a particularly salient form of anthropocentric ideology in mainstream AI ethics, and to resist its use to politically justify and institutionalize those oppressive practices. This approach does not imply a comparison between the (potential) experiences of artificial agents and the multiple intersecting forms of discrimination and oppression faced by black and brown people, women and LGBTQIA+ people, and other marginalized groups under white supremacy, cisheteronormative patriarchy, and other entrenched systems of power. Nocella says, "All suffering is different and is based on individual experience even if the oppressive tactic is the same" (Nocella, 2012, p. 147). Our goal is not to speak on behalf of or "liberate" robots, nor to merely appropriate the language and culture of activist movements. Instead, our goal is to contribute to the struggle against all forms of oppression by examining one manifestation of a tactic that impacts the operation and public participation of both humans and nonhumans alike; namely, the political classification of agents by *type*, and the systematic privileging of some agential types relative to others (Benjamin, 2016; Braun, 2014; Reardon, 2009).

We adapt the term "human supremacy" from the context of animal and environmental advocacy to the field of AI in order to name a nefarious mode of classification politics that places the "human" as the locus of systemic privilege. Used in this way, the term retains much of its original meaning. Nevertheless, the critique of human supremacy in AI presents several unique challenges that distinguish it from the animal advocacy case. One important difference is that while anthropocentrism in environmental policy can be subtle and may require critical interpretive efforts to "recognize" (Lupinacci, 2015), in AI human supremacy is often *overt*, with the "human" presented as the explicit basis for political alliance, as we will see in Bryson's view in the next section. To this extent, the term "human supremacy" functions less as an accusation of covert oppressive behavior by analogy to racial oppression, and more as an accurate description of an ideological position framed in its proponent's own terms.

---

expansion of for-profit concentration camps at the border that have kept innocent people in terrible conditions and have separated thousands of children from their parents.

[6] The term "AI" and "robot" are used to include all technologies addressed under the labels artificial intelligence, machine learning, robotics, autonomous vehicles and drones, autonomous weapons systems and surveillance equipment, IoT and "smart" appliances, social media bots and other artificial software agents (anthropomorphic or not), various expert systems and efficient database management architectures, and nearby technologies. The terms for identifying and distinguishing between artificial agents playing various roles remains largely unsettled. Occasionally the term "system" or "machine" will be used as inclusive of humans and nonhumans.

Before describing Bryson's view in more detail, we should discuss why human-centered politics is so broadly welcomed in AI ethics, despite the dismal status of the human in the political climate of the Anthropocene (Ellis, 2015; D. J. Haraway, 2016; Lewis & Maslin, 2015). The literature discusses two justifications which have, on their surface, relatively little to do with each other: the existing international policy framework of human rights (Latonero, 2018; Risse, 2018), and the presumed metaphysical non-agency of artifacts like machines and pieces of software (Boden et al., 2017; Fossa, 2018). These justifications will be addressed in later sections through the lens of Braidotti's critical posthumanism and Mills' critique of ideal theory respectively. However, one version of the second justification should be addressed before moving from the animal ethics literature.

It is commonly argued that AI is neither biological nor alive, so cannot suffer like animals and other living creatures, and therefore cannot participate in a moral community in the relevant way to deserve moral consideration. If animal advocacy is primarily motivated by animal suffering, and robots cannot suffer, this would seem to undermine the possibility that robots could stand in need of the sort of political activism seen in the animal advocacy movement. Such arguments are not convincing on several grounds. Environmentalists since Aldo Leopold's *Land Ethic* (1949) have emphasized the value of nonliving systems like the soil, water, and air that do not "suffer" in the experiential sense of animals with a nervous system, but which nevertheless are vital for the integrity of ecological communities, and so might play a focal role in our norms and practices (Konopka, 2013). Kate Darling (2016) notes that while the philosophical and ethical discussion of animal rights revolves around issues like pain and consciousness, "our laws indicate that these concerns are secondary when it comes to legal protections" (Darling, 2016, p. 17). Instead, Darling argues that laws tend to follow public attitudes towards animals that do not depend on biological differences, as with laws in the US that protect horses but not cows from being killed and eaten, despite few biological differences that could justify this practice. Several scholars have noted how the discourses on conscious experiences in AI privilege a Western European and predominantly Christian perspective on artifacts and their relationship to nature and society—a perspective that is not universally shared (Gunkel, 2018a; Jones, 2015; Williams, 2019). There are also the ongoing discriminatory practices in which the appeal to biology is treated as scientific justification for institutional oppression (Appiah, 2018). Taken together, these considerations suggest that biological factors should not be treated as *prima facie* justification for the exclusion of artificial agents from the moral community.

Nevertheless, the literature on "robot rights" is overwhelmingly preoccupied with whether robots have experiential or conscious states sufficiently "like ours" to warrant social status and legal recognition (Danaher, 2019; Gunkel, 2018b; Schwitzgebel & Garza, 2015)[7]. The hypercritical focus on the machine's experience (or lack thereof)

[7] These concerns predate Turing's (1950) proposed "imitation game," which can be read as an attempt to redirect questions away from the machine's "experience" and towards the actual conditions of its social performances, including our reactions to them (Estrada, 2018; Hayles, 2008). Notice how the systematic comparison between human and machine experiences suggested by the popular reading of Turing's test runs afoul of Nocella's suggestion to avoid comparing experiences. This is not to suggest that Turing's test is morally wrong, but to

points to another important distinction between animals and AI: artificial agents are already visibly engaged in a variety of human sociopolitical contexts, and there is some reasonable expectation that their capacities will improve over relatively short time scales. To pick a benign example, while it is extremely unlikely that the next 1000 generations of domesticated cat will develop an affinity for poetry, there are currently bots generating poems on Twitter with hundreds of followers (Oliveira, 2017). Such systems are already common enough that scholars have begun to discuss the aesthetics of AI art (Hertzmann, 2019). The overlap in the sociopolitical circumstances of human and artificial agents is not predicated on some shared biological or ecological background, nor on shared experiences or conscious states, but more concretely on the material and institutional realities within which human and nonhuman agents "share existence" (Latour, 2003). Just as the shared material realities of oppression provide a framework for collaboration among resistance movements addressing both human and nonhuman animal interests—despite important differences in the history and experiences motivating this work—this very same framework of resistance provides resources to critique and resist biocentrism and anthropocentrism in the discourse around AI and artificial agency, despite a strong tradition emphasizing the differences between biological and artificial agents.

The possibility for radical near-term change in the agential capacities of AI suggests its sociopolitical status will likewise remain unsettled. Asaro (2006) notes, "At some point in the future, robots might simply demand their rights" (Asaro, 2006, p. 12). However, such a future would require not just a change in the capacities of artificial agents, but also an ethical and interpretive change in society's capacities to recognize and respect such demands as legitimate political acts. The demand for cultural changes in our attitudes towards artificial agents echoes Turing's plea of "fair play for machines" (Estrada, 2018a; Turing, 1947). If the kind of critical self-reflection required to advocate on behalf of machines was available to Turing, surely it is available to us as well.

Human supremacy in Bryson's ethics

Joanna Bryson is a Reader and Senior Research Fellow in the Department of Computer Science at the University of Bath. Outside her work in AI ethics, Bryson's research focuses on computer models of economic and cooperative behaviors, with strong influences from cognitive and evolutionary psychology. Although this research somewhat overlaps with her work in AI ethics (Bryson, 2015a), it is beyond the scope of this paper to address Bryson's full research project. Instead, the goal of this section is to present a critical reading of the article "Robots Should Be Slaves" **(RSBS)** (Bryson, 2010) and other papers that shed light on its motivation, perspective, and influence on Bryson's efforts as a high level policy consultant in AI. Despite its theoretical problems, Bryson's work has the virtue of clarity and precision (sometimes unwittingly) in the articulation of her views, and so serves as a model case for studying human supremacist ideology in AI.

recognize that it creates social circumstances that are especially hostile to the possibility of recognizing and respecting artificial agents (Hayes & Ford, 1995).

Written nearly a decade before RSBS, Bryson's earliest contribution to AI ethics scholarship is the coauthored conference paper, "Just Another Artifact: Ethics and the Empirical Experience of AI" (Bryson & Kime, 1998) which lays out many of the elements of her considered view. Bryson & Kime's explicit motivation in this paper is to address certain "exaggerated fears" (p. 1) from Vernor Vinge and other early proponents of the Singularity hypothesis: that computers might surpass human intelligence and take over the world (Vinge, 1993). Bryson & Kime argue that these misplaced fears come from an "over-identification with machines," a mistake that is "symptomatic of a larger problem—a general confusion about the nature of humanity and the role of ethics in society." (p. 1) What is the nature of humanity and the role of ethics in society? The authors claim that "ethics has evolved as a mechanism of human social cohesion, without which society disintegrates" (p. 2). The primary mechanism driving social cohesion is empathy: "we care for people or objects that we would *feel badly for* if they were hurt or damaged" (*ibid*, original emphasis). This feeling of empathy in turn creates a sense of *identification* with those we empathize with. The relative strength of this identification generates an individual's hierarchy of ethical obligation, "with ourselves and our families tending to be at the top, followed by our neighbours and other people with whom we acknowledge commonality" (*ibid*). Bryson & Kime argue that this general picture suggests that, "self-interest is the root of our ethics" (p. 4).

In the case of the Singularity theorists, Bryson & Kime argue that the mechanism of social identification has been misapplied to machines. They explain this confusion by appeal to our tendency to distinguish ourselves from animals—itself grounded in the evolutionary drive to empathetic social cohesion. They claim, "To form a human society, one needs to value the lives of humans in the community over the lives of other animals" (*ibid*). They argue that over-identification with machines "lead[s] to an undervaluing of the emotional and aesthetic in our society. Consequences include an unhealthy neglect and denial of emotional experiences" (*ibid).* They identify two specific dangers of over-identification, that, "we may believe the machine to be a participant in our society, which might seriously confuse our understanding of them," and "we may overvalue the machine when making our own ethical judgments and balancing our own obligations" (*ibid*). They dismiss any view that values AI, "to the exclusion of our own existence" as "nihilism" (p. 6). The most substantial citation offered for these claims is (Lakoff & Johnson, 1980).

To be clear, there are many reasons for finding these views unsatisfying as an ethical framework. The direct line drawn between evolutionary psychology and ethical obligation is theoretically implausible (Street, 2006). The relationship proposed between empathy and social identification is, at best, oversimplified (Jenkins, 2014). The view that AI is only dangerous through its misuse or abuse by humans is known in the philosophy of technology literature as "technological neutrality" or "instrumentalism," (Kaplan, 2009), a problematic view as a policy position in tech development (Koops, 2006; Reed, 2007; Winner, 1980). Attributing to her opponents' "unhealthy" psychological disorders is a questionable rhetorical tactic that was rightly edited out of the 2011 version. However, it is not our goal to present a scholarly critique of a conference paper from more than twenty years ago. Instead, our goal is to trace the development of a view that results in the explicit endorsement of *slavery* as a political framework for managing robots. For our purposes, the most relevant features of

Bryson's position in this paper are the claims that identification drives moral obligations, and that identifying with AI is a *mistake*.

Neither claim is compelling. On purely psychological grounds, social identification is unlikely to build models that are consistent enough to serve as a basis for moral reasoning. Jenkins (2014) introduces the psychology of social identity by explaining,

> "… our classificatory models of self and others are multidimensional, unlikely to be internally consistent, and may not easily map onto each other. Hierarchies of collective identification may conflict with hierarchies of individual identification, which means that the following might make complete interactional sense: I hate all As; you are an A; but you are my friend. Taken together, these points suggest that categorical imperatives are unlikely to be a sufficient guide on their own, and that the ability to discriminate between others in subtle and fine-grained ways is an everyday necessity." (Jenkins, 2014, p. 6)

Bryson's later work emphasizes that moral systems should be "coherent" (Bryson, 2018b, p. 202) If identification does not produce consistent moral frameworks, it seems highly unlikely that "over-identification" (with machines or anything else) is a serious threat to the social order. Suggesting that AI policy should center on cultivating the appropriate social identification practices seems impractical, to say nothing of its ethics.

For the sake of argument, however, suppose we accept Bryson's first claim. If identification is the root of obligation, then the psychological fact that we identify with machines would suggest some obligations to those machines. What justifies the claim that such identification is inappropriate or mistaken? Bryson & Kime (1998) recognize that ethical systems can be "somewhat arbitrary," and that in novel circumstances (as with AI), we are "to some extent free to create a new ethical standard" (p. 5). So why not take our identification with machines as evidence of a new set of moral obligations? The authors defend their judgment with two responses, one they describe as "technical", the other "ethical" (p. 3). Their technical response recognizes that people tend to ascribe capabilities to machines that those machines don't have. Their ethical response recognizes that some people might already believe that some computer programs are more valuable than a human life, but that we face a similar challenge with other artifacts like "fine art and political institutions" (p. 3).

Neither response addresses this issue at stake, which is how we decide which forms of identification (and which identities) are appropriate or inappropriate. As to the technical response: we identify not just with other people, but with sports teams and brand names and superheroes and all manner of things. The fact that we make errors about the capacities of these entities says very little about whether our identification with them is appropriate or inappropriate. When fans of an underperforming sports team are unrealistically optimistic about their performance in tonight's game, this is not evidence that their identification with the team is mistaken, inappropriate, or symptomatic of deeper psychological or conceptual problems. The ethics of identification are not settled by the accuracy of the predictions it generates.

This technical response is particularly confusing given that their proposal contains much better resources for addressing the concern: specifically, the evolutionary drive to "social cohesion." On this supposedly ethics-grounding biological

imperative, the precise nature of the ethical system doesn't matter so much as its overall impact on social stability and (ultimately) the reproductive success of the species. This would seem to make the issue an open empirical question: does empathizing with machines make for a more stable social order? Or perhaps better as an engineering and design question: how do we design more stable social systems through the natural empathy people have towards machines? By insisting that the identification with machines is a *conceptual mistake*, Bryson & Kime cut off these possibilities and effectively limit the ethical discourse to controlling the frequency and impact of these "anthropomorphic fallacies" (p. 7). In this spirit they claim, "The issue of forming identity is now more than ever an issue for public education," suggesting the need for institutionalized policies that control how social identities are formed and who we identify with.

Bryson & Kime's second "ethical" response reveals an important assumption in Bryson's ethical perspective: that the evolutionary dynamics of obligation are *zero sum*, and that developing new obligations towards AI would entail fewer obligations to humans, animals, and society generally. The risk is not simply that we identify with AI, but that we identify with AI *at the expense of identifying with humans*; if obligation is zero sum, these identities are necessarily in competition. Since they assume that obligations to other humans is vital for social cohesion, then over-identification with AI is not merely inappropriate; it presents a clear threat to the social order. They recognize that this threat is not unique to AI, and point to the resources used to maintain the Mona Lisa that could be used instead for people in need. Restating this argument, Bryson & Kime are claiming that *great art threatens social cohesion* (and our evolutionary success!) by potentially generating more empathy for art than we have for other people. Their criticism of AI is that it might pose the same threat to social cohesion as posed by great art. One might have thought that art provides a clear example in which identification and obligation are *not* zero-sum, where we might be a more stable, cohesive, empathetic society because of the resources we invest in public art. But for Bryson & Kime, producing great art is a social *liability*, a risk we accept, like car accidents, because of the pleasure and convenience those artifacts bring us.

Despite their weaknesses, these positions have a strong influence on Bryson's later work. RSBS is published as a book chapter in 2010 after being solicited for a conference on "Artificial Companions in Society" at Oxford in 2007. Its publication coincides with a burst of papers and conference presentations with titles like "Building Persons is a Choice" (Bryson, 2009); "Why robot nannies probably won't do much psychological damage" (Bryson, 2010b); and "AI/Robots should not be considered moral agents" (Bryson, 2011). These papers all expand on the themes developed in "Just Another Artifact," highlighting the ethical challenge of over-identification with AI and the conceptual mistakes and dangers it invites. Clearly, Bryson's immediate goal in this collection of work is not to subjugate robots, but to correct what she sees as the conceptual confusion generated by an inappropriate identification with machines. In 2010, Bryson participated in a retreat that produced a set of "Principles of Robotics" (Boden et al., 2017) that reinforces these themes. Of the five rules laid out in the document, the first four simply state *what robots are*: "1: Robots are multi-use tools…, 2: Humans, not robots, are responsible agents…, 3: Robots are products…, 4: Robots are manufactured artifacts…" (Boden et al., 2017, p. 125ff). As a policy proposal, the

document recommends the development of industrial identification practices that clearly distinguish between the capacities of robots and humans. These principles give the myopic impression that the primary ethical risk presented by AI is a metaphysical and ontological confusion over their agential status.

The theoretical grounding for RSBS is now brought into focus. By asserting "robots should be slaves" (Bryson, 2010a), Bryson takes for granted not only that humans should not be treated as slaves, but also that *no one would identify or empathize with slaves*. Calling *robots* slaves draws attention to this strong categorical distinction between humans and robots, and challenges the excessive empathy with AI that she takes to be the central risk at stake. In RSBS, Bryson repeats the claim of that "our identity confusion results in somewhat arbitrary assignments of empathy" (p. 4) and lists a set of costs for both individuals and institutions associated with the over-identification with AI (p. 5). She calls "being too generous with personhood" a "moral hazard" (p. 7). She rehearses the zero-sum reasoning[8], arguing that "humans have only a finite amount of time and attention for forming social relationships" (p. 5). While Bryson recognizes that the costs of identification "could be negative," she doesn't spend much time discussing the social benefits of identifying with machines, or how to design machines that maximize these benefits.

Instead, she uses her list of perceived costs to motivate what she calls the "correct metaphor" for robotics: that "robots should be servants you own" (p. 3). She says, "communicating the model of robot-as-slave is the best way both to get full utility from these devices and to avoid the moral hazards" (p. 8). RSBS lists "the fundamental claims of the paper" as:

> "1. Having servants is good and useful, provided no one is dehumanized."
> 2. A robot can be a servant without being a person.
> 3. It is right and natural for people to own robots.
> 4. It would be wrong to let people think that robots are persons." (p. 3)

As the list suggests, Bryson's concern for dehumanization is mostly an afterthought. She says, "Surely dehumanization is only wrong when it's applied to someone who really is human?" (p. 2) Bryson goes on to briefly discuss the history of domestic labor in British villages from 1574-1821 in a positive light, claiming that roughly 30% of households employed servants. She justifies this practice by appeal to the inadequacies of an unpaid gendered division of labor, saying, "Where wives and other kin were not available to devote their full time to these tasks, outside employees were essential" (p. 8)[9]. Bryson reflects favorably on the current market for domestic labor, but argues "the most difficult thing with human servants is of course the fact that they really are humans, with their own goals, desires, interests, and expectations which they deserve to be able to pursue" (p. 9).

---

[8] In (Bryson et al 2017), Bryson et al recognize the potential problem with this zero-sum reasoning, but they appeal to it anyway, saying "While not always a zero-sum game, sometimes extending the class of legal persons can come at the expense of the interests of those already within it. In the past, creating new legal persons has sometimes lead to asymmetries and corruptions such as entities that are accountable but unfunded, or fully-financed but unaccountable."

[9] If they were essential, what did the other 70% do?

On the other hand, because robots "are wholly owned and designed by us" (p. 9), they cannot be frustrated, exploited, or made to suffer unless we deliberately design them with these capacities. So long as we aren't anthropomorphizing robots in ways that causes confusion or excessive empathy, she says "owners should not have ethical obligations to robots… beyond those that society defines as common sense and decency, and would apply to any artifact" (p. 10). Bryson admits we have ethical obligations *concerning* robots, about their safe operation and so on, but we have no obligations to the robots themselves; destroying a robot is ethically equivalent to the destruction of any property. In one of the more frustrating passages (p. 8), Bryson suggests that people who aren't comfortable with the metaphor of slavery might instead adopt the perspective of extended mind theory (Clark & Chalmers, 1998), where our tools are understood as extensions of our own capacities. Bryson doesn't consider that the extended mind theory encourages us to strongly identify with our machines (Ahuvia, 2005), or why this might be inconsistent with her proposal for robots-as-slaves.

Obviously, the appeal to the institution of slavery as "good and useful, … right and natural" is profoundly insensitive and simply in poor taste. It also highlights the deep theoretical failure in Bryson's ethics. Just as with the *Mechanix Illustrated* comic from 1965, Bryson takes for granted that the public would identify with slave owners rather than slaves, and with the 30% of the British who hired domestic servants, rather than the 70% from whom they were hired. These assumptions speak volumes on Bryson's perception of her own social status and ethical obligations. More importantly, they speak to the substantial challenges involved in grounding ethical policy in social identity construction, challenges Bryson clearly did not anticipate when writing this essay. Although the essay makes token gestures to recognize the cruelty of the historical institution of racialized slavery, she takes no effort to consider how the metaphor of slavery might be interpreted by those who don't immediately identify—that is, empathize—with slaveholders over slaves, or the moral hazards involved with giving any defense of slavery as a political institution. Fundamentally, Bryson does not think the problem with slavery was the ideology of domination and oppression it represents; the problem with slavery was that we were enslaving the wrong things! Thankfully, experts like Bryson are here to make sure we're practicing slavery correctly.

RSBS has received substantial scholarly attention in the AI ethics literature (Agar, 2019; Coeckelbergh, 2015; Frank & Nyholm, 2017; Gunkel, 2015, 2018b; Musiał, 2017; Neely, 2014; Prescott, 2017; Rainey, 2016; van Wynsberghe & Robbins, 2019). While much of this literature is critical of Bryson's hasty and insensitive language, few engage Bryson's theoretical approach from the perspective of standpoint epistemology or critical race theory. As a consequence, the primary lesson Bryson has drawn from this criticism is that "you cannot use the term 'slave' without invoking its human history" (Bryson, 2015b). In other words, she thinks criticisms are reacting to her choice of language, not to the oppressive ideology that language articulates, or the casual neglect it shows for those who don't share her identities or perspective.

So, despite these scholarly criticisms, Bryson's work continues to develop the central perspective in RSBS. In "Of, for, and by the people: the legal lacuna of synthetic persons" (Bryson et al., 2017), Bryson claims that "the basic purposes of human legal systems" include a principle that, "Should equally weighty moral rights of two types of entity conflict, legal systems should give preference to the moral rights held by human

beings." (Bryson et al., 2017, p. 283) Notice the tacit assumption that part of the basic purpose of human legal systems is *to sort entities by type*. Lest there remains any ambiguity in Bryson's intentions, she describes her position as, "an uncontroversially light thumb on the scale in favor of human interests. Yes, this is speciesism" (*ibid*). In "Patiency is not a virtue: the design of intelligent systems and systems of ethics", Bryson argues that making robots deserving to be moral patients "could in itself be construed as an immoral act" (Bryson, 2018b, p. 16). In "No One Should Trust AI" (Bryson, 2018a), Bryson argues that trust is a relationship between peers, and since we aren't peers with AI, "no one actually *can* trust AI" (emphasis original). In these papers, one of Bryson's explicit concerns is that malicious corporate actors might anthropomorphize machines in order to exploit our psychological biases and legal loopholes. While the language of slavery is absent from these essays, her solutions nevertheless involve policy that imposes a strict hierarchy in which machines are categorically subordinate to human interests. Since humans and machines are not distinguished by species, describing this view as "speciesism" is inaccurate. Bryson defense of policy that establishes an explicit social hierarchy modelled on slavery that classifying and systematically privileges humans deserves the name human supremacy.

Dismissing these views as eccentric or theoretically untethered ignores the rise to prominence as a speaker, consultant, and high-level policy expert that Bryson has enjoyed in AI ethics on the strength of this work. In 2017, Bryson was quoted calling the popular humanoid robot Sophia's award of Saudi citizenship "bullshit" (Vincent, 2017). The comment earned her a public debate with Sophia creator David Hanson at CogX 2018 (Estrada, 2018b). In 2019, Bryson was selected to sit on the controversial Advanced Technology External Advisory Council at Google (Johnson, 2019), a body she continued to defend after activist protest shut the program down (Bryson, 2019). To some extent, Bryson was in the right place when the AI boom hit to find success in the developing field of AI ethics. At the same time, her success can be at least partly attributed to an industry and regulatory climate that was particularly receptive to the vision of human-centered ethics her work developed. To address the broader milieu in which Bryson's work finds success, we turn next to the work of Rosi Braidotti and Charles Mills.

Reactionary posthumanism and ideal theory

If, as Bryson suggests, humanity is in the grips of an identity crisis, then Braidotti's framework of "the posthuman" (Braidotti, 2013) may help diagnose the problem. For Braidotti, posthumanism marks an historical condition characterized not only by a "crisis of Humanism," but also the active exploration of "alternative ways of conceptualizing the human subject" (p. 37). Braidotti identifies three strands of posthuman thought that trace out different responses to our posthuman condition: one, a *critical* posthumanism informed by anti-humanist philosophies of subjectivity from Braidotti's own scholarly tradition (Braidotti, 1994, 2002; Foucault, 1977); second, an *analytic* posthumanism that derives from explorations of the human in science and technology studies (Roden, 2014; Verbeek, 2005, 2011); and finally, a *reactionary* posthumanism for whom "the posthuman condition can be solved by restoring a humanist vision of the subject"

(Braidotti, 2013, p. 39). Braidotti's discussion of reactionary posthumanism is most relevant for our treatment of Bryson's views.

Braidotti associates reactionary posthumanism with Martha Nussbaum (1998, 2010) who, Braidotti argues, "defends the need for universal humanistic values as a remedy for the fragmentation and relativistic drift of our times" (Braidotti, 2013, p. 39). For Nussbaum, this fragmentation is produced by the socioeconomic condition of globalization, which threatens humanity through the reactionary "plagues" (*ibid*) of ethnocentrism and xenophobic nationalism. According to Braidotti, Nussbaum believes that the solution to these threats is a cosmopolitan universalism informed by classical humanist ideals. Braidotti says that for Nussbaum, "abstract universalism is the only stance that is capable of providing solid foundations for moral values such as compassion and respect for others" (*ibid).* Nussbaum acknowledges the problematic historical use of humanist ideals as a discriminatory or exclusive practice, and she responds to these past failures with a call for a neo-humanism that centers the subjectivity of experience. While Braidotti praises this move, she argues that Nussbaum, "reattaches [subjectivity] to a universalistic belief in individualism, fixed identities, steady locations and moral ties that bind" (*ibid).* Because of this "disembedded universalism, Nussbaum ends up being paradoxically parochial in her vision of what counts as the human… leaving no room for experimenting with new models of the self" (*ibid*).

For example, Braidotti describes Nussbaum's defense of a liberal education (Nussbaum, 2010) as "elitist and nostalgic" (Braidotti, 2013, p. 173), noting that by the time of its publication, the university had already been refigured in the market economy as a corporate structure (p. 150). Braidotti's point is not simply to disagree with Nussbaum about the value of a liberal education, but rather to recognize that the liberal ideals which ground Nussbaum's defense are out of touch with the material and institutional realities which benefit from that defense. If universities are managed like for-profit corporate chains, this muddies the narrative of the liberal ideal that the university supposedly represents. Similarly, if the rhetoric of universalist humanism is used to protect narrow and exclusive practices, it undermines the appeal of those humanist ideals. The point is not that the humanist ideals—for instance, the framework of human rights in international justice—are necessarily bad as a statement of absolute morality, but rather that the ideals are easily used to defend the very practices that subvert them. Similarly, when humanist ideals are used to justify an exclusive attitude towards artificial agency its apparently inclusive framing is undermined.

Charles Mills' (2005) critique of "ideal theory" as ideology provides conceptual tools for thinking through this potentially confusing discursive situation. For Mills, "ideal theory" describes not just the appeal to idealizations, which to some extent cannot be helped in theoretical discourse. Rather, ideal theory describes "the reliance on idealization to the exclusion, or at least marginalization, of the actual" (p. 168). For instance, ideal theory might concern itself with how an ideal society would structure its basic institutions, rather than addressing the social circumstances in which its actual institutions operate. Mills claims that ideal theory will typically employ assumptions that idealize human capacities, social institutions, and social ontology in ways that "abstract away from relations of structural domination, exploitation, coercion, and oppression, which in reality, of course, will profoundly shape the ontology of those same individuals"

(p. 168). Mills says, "It is obvious that ideal theory can only serve the interests of the privileged who, in addition—precisely because of that privilege (as bourgeois white males)—have an experience that comes closest to that ideal, and so experience the least cognitive dissonance between it and reality" (p. 172).

Restating the critique of Nussbaum in Mills' terms, Braidotti is accusing Nussbaum of embracing universalist humanism as an idealization to the exclusion of the actual. It is beyond the scope of this paper to assess whether Braidotti's criticisms are fair to Nussbaum's actual views. What matters for our purposes is Braidotti's analysis of how a reactionary embrace of humanist ideals in a posthuman context can suffer from the ideology of ideal theory. The institutional realities of the corporate influence on the university system have real implications for our discussion of the value of higher education, and this influence can't be dismissed or marginalized by appeal to classical humanist ideals. Mills explains that by, "abstracting away from realities crucial to our comprehension of the actual workings of injustice in human interactions and social institutions" (p. 170), the appeal to ideal theory effectively guarantees that those ideals will never be achieved.

Together, Braidotti and Mills help articulate the critical failures in Bryson's ethics beyond the mere insensitivity of her language. Like Nussbaum, Bryson responds to the crisis of humanism by asserting nostalgic, elitist idealizations of social institutions and social ontology, such as "servants are useful and good," or "no one should trust AI." Bryson presents herself as speaking on behalf of humanity's interest, when in fact her proposal showcases a narrow and privileged perspective that inadvertently alienates those who don't already share it. In so doing, Bryson neglects the structures of power and domination within which humans and nonhumans share existence.

Human supremacy as posthuman risk

Consider, for instance, the startup Kiwibots, which offers a food delivery service using small robots in the Bay area. Kiwibots set itself apart from similar services in deciding against using AI software to control their bots. Instead, as reported in the Chronicle (C. Said, 2019), they farm out the control task out to operators in Colombia who use GPS to direct the bot to its destination. The Colombian operators are paid less than $2 an hour, which the Chronicle says is more than the local minimum wage. Kiwibots provides an interesting case at the intersection of automation, teleoperation, and the global labor market. It also provides a confounding case for idealizing proposals like Bryson's that insist on a strict dichotomy between humans and machines. Social policy treating this robot as a *slave* would be indirectly treating another human as a slave, with many of the same structures of exploitation and oppression the term invites. Though Bryson explicitly rejects forms of servitude that lead to dehumanizing *people*, in this case we can't easily distinguish between the human and the robot, so we can't tell if our dehumanizing behavior is "appropriate" or not.

Such cases are by not unique. The New York Times reported in May that Google's automated assistant Duplex transfers around 25% of calls to a human operator (Chen & Metz, 2019). After Duplex's controversial debut, there have been several proposals for laws that would require such automated services to identify

themselves as bots. These "bot disclosure" laws have faced objections from civil rights groups like the EFF, who worry that an "across the board bot-labeling mandate would sweep up all bots," including those being used for protected speech (EFF, 2018). The relative difficulty in distinguishing between human and machine behavior already means that solving CAPTCHAs (Von Ahn, Blum, Hopper, & Langford, 2003) is part of the daily life of a digital citizen. It's worth recognizing that CAPTCHAs were originally developed not only to filter out bots, but also as a source of free labor to crowdsource difficult problems in early machine learning (Lung, 2012). This highlights the deep connections between bot identification, the exploitation of labor, and the industry's penchant for the systematic classification of agents by type (Stieglitz, Brachten, Ross, & Jung, 2017; Suárez-Serrato, Velázquez Richards, & Yazdani, 2018).

What's the alternative to the ideal theory of reactionary posthumanism? Following Mills, Serene Khader advocates for a *nonideal universalism* that emphasizes the nonideal, unjust conditions of political action (Khader, 2018). Khader says, "One defect of ideal theories… is their tendency to redirect our evaluative gazes to the wrong normative phenomena" (p. 36). By redirecting that gaze towards the actual conditions of injustice, Khader argues that we will be in a better position to address those injustices. While Khader's nonideal universalism is developed in the context of transnational feminism, we might adapt the approach to an AI ethics context by recognizing the nonideal and unjust conditions in which both human and nonhuman agents operate. A nonideal approach to AI ethics would highlight how overlapping structures of institutional oppression situate robots as both agents *and* targets of power—as agents whose identity must be made available for inspection, public scrutiny, and abuse (Brscić, Kidokoro, Suehiro, & Kanda, 2015; Romero, 2018; Salvini et al., 2010; Smith & Zeller, 2017). Human supremacist rhetoric reinforces these conditions, both drawing on and reinforcing the posture of other oppressive ideologies which control the social order through the classification and systematic privileging of agents by type. Advocating on behalf of robots as social participants deserving of minimal respect and dignity in virtue of that participation can encourage nothing but a more respectful, participatory, and dignified social order.

Acknowledgements

Portions of this work were presented at the Gender, Bodies, and Technology 2019 Conference in Blacksburg, VA, and the Computer Ethics—Philosophical Enquiry 2019 in Norfolk, VA. Thanks to the support and wisdom of Anna Gollub, Patrick O'Donnell, Anna Lauren Hoffmann, Os Keyes, Ruha Benjamin, Eric Schwitzgebel, David Gunkel, Robin Zebrowski, Ashley Shew, Jonathan Flowers, Damien Williams, Joshua Earle, Justin Remhof, Dylan Wittkower, Brandon Stanford, Kelly Kitchens, Conner Downey, Austin Landini, Thomas Shea, Todd Kukla, Shawn and Mindy Dingle, the organizers and participants at GBT'19 and CEPE'19, and my wonderful students and colleagues at NJIT and CTY Princeton.

References

Agar, N. (2019). How to Treat Machines that Might Have Minds. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00357-8

Ahuvia, A. C. (2005). Beyond the Extended Self: Loved Objects and Consumers' Identity Narratives. *Journal of Consumer Research*, *32*(1), 171–184. https://doi.org/10.1086/429607

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*, 2016.

Appiah, K. A. (2018). *The lies that bind: Rethinking identity*. Profile Books.

Armstrong, P. (2002). The Postcolonial Animal. *Society & Animals*, *10*(4), 413–419. https://doi.org/10.1163/156853002320936890

Asaro, P. (2006). What should we want from a robot ethic. *International Review of Information Ethics*, *6*(12), 9–16.

Asaro, P. (2016). Hands up, don't shoot!: HRI and the automation of police use of force. *Journal of Human-Robot Interaction*, *5*(3), 55–69.

Bardzell, J., & Bardzell, S. (2015). Humanistic Hci. *Synthesis Lectures on Human-Centered Informatics*, *8*(4), 1–185.

Belcourt, B.-R. (2015). Animal bodies, colonial subjects:(Re) locating animality in decolonial thought. *Societies*, *5*(1), 1–11.

Benjamin, R. (2016). Catching Our Breath: Critical Race STS and the Carceral Imagination. *Engaging Science, Technology, and Society*, *2*(0), 145–156. https://doi.org/10.17351/ests2016.70

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Retrieved from https://books.google.com/books?id=G6-hDwAAQBAJ

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., … Winfield, A. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science*, *29*(2), 124–129. https://doi.org/10.1080/09540091.2016.1271400

Braidotti, R. (1994). *Nomadic subjects: Embodiment and sexual difference in contemporary feminist theory*. Columbia University Press.

Braidotti, R. (2002). *Metamorphoses: Towards a materialist theory of becoming*. Cambridge: Polity Press.

Braidotti, R. (2013). The Posthuman. *Cambridge: Polity*.

Braun, L. (2014). *Breathing race into the machine: The surprising career of the spirometer from plantation to genetics*. U of Minnesota Press.

Brscić, D., Kidokoro, H., Suehiro, Y., & Kanda, T. (2015). Escaping from children's abuse of social robots. *Proceedings of the Tenth Annual Acm/Ieee International Conference on Human-Robot Interaction*, 59–66. ACM.

Bryson, J. (2009). Building persons is a choice. *Erwägen Wissen Ethik*, *20*(2), 195–197.

Bryson, J. (2010a). Robots should be slaves. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, 63–74.

Bryson, J. (2010b). Why robot nannies probably won't do much psychological damage. *Interaction Studies*, *11*(2), 196–200. https://doi.org/10.1075/is.11.2.03bry

Bryson, J. (2011). AI robots should not be considered moral agents. In N. Berlatsky (Ed.), *Artificial Intelligence*. Greenhaven Press.

Bryson, J. (2015a). Artificial Intelligence and Pro-Social Behaviour. In C. Misselhorn (Ed.), *Collective Agency and Cooperation in Natural and Artificial Systems* (pp. 281–306). https://doi.org/10.1007/978-3-319-15515-9_15

Bryson, J. (2015b, August 4). Clones should NOT be slaves. Retrieved August 5, 2019, from Adventures in NI website: https://joanna-bryson.blogspot.com/2015/10/clones-should-not-be-slaves.html

Bryson, J. (2018a). AI & Global Governance: No One Should Trust AI. *Centre for Policy Research at United Nations University*. Retrieved from https://cpr.unu.edu/ai-global-governance-no-one-should-trust-ai.html

Bryson, J. (2018b). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15–26.

Bryson, J. (2019, April 7). What we lost when we lost Google ATEAC. Retrieved August 11, 2019, from Adventures in NI website: https://joanna-bryson.blogspot.com/2019/04/what-we-lost-when-we-lost-google-ateac.html

Bryson, J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*(3), 273–291.

Bryson, J., & Kime, P. (1998). Just another artifact: Ethics and the empirical experience of AI. *Fifteenth International Congress on Cybernetics*, 385–390.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Chen, B. X., & Metz, C. (2019, May 22). Google's Duplex Uses A.I. to Mimic Humans (Sometimes). *The New York Times*. Retrieved from https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Coeckelbergh, M. (2015). The tragedy of the master: Automation, vulnerability, and distance. *Ethics and Information Technology*, *17*(3), 219–229.

Crist, E. (2017). The affliction of human supremacy. *The Ecological Citizen*, *1*, 61–4.

Danaher, J. (2019). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, 1–27.

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. Froomkin, & I. Kerr, *Robot Law* (pp. 213–232). https://doi.org/10.4337/9781783476732.00017

Deckha, M. (2012). Toward a postcolonial, posthumanist feminist theory: Centralizing race and culture in feminist work on nonhuman animals. *Hypatia*, *27*(3), 527–545.

EFF Letter Opposing California Bot Disclosure Bill, SB 1001—First Amendment Concerns. (2018, May 21). Retrieved August 12, 2019, from Electronic Frontier Foundation website: https://www.eff.org/document/eff-letter-opposing-california-bot-disclosure-bill-sb-1001-first-amendment-concerns

Ellis, E. C. (2015). Ecology in an anthropogenic biosphere. *Ecological Monographs*, *85*(3), 287–331. https://doi.org/10.1890/14-2274.1

Estrada, D. (2018a). Value alignment, fair play, and the rights of service robots. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 102–107. ACM.

Estrada, D. (2018b, June 18). Sophia and her critics. Retrieved August 11, 2019, from Medium website: https://medium.com/@eripsa/sophia-and-her-critics-5bd22d859b9c

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*. https://doi.org/10.1007/s11023-018-9482-5

Fossa, F. (2018). Artificial moral agents: Moral mentors or sensible tools? *Ethics and Information Technology*, *20*(2), 115–126.

Foucault, M. (1977). *Discipline and Punish*. New York: Pantheon Books.

Frank, L., & Nyholm, S. (2017). Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial Intelligence and Law*, *25*(3), 305–323. https://doi.org/10.1007/s10506-017-9212-y

Frye, M. (1983). Oppression. The politics of reality: Essays in feminist theory. *Reprinted in V. Taylor, N. Whittier, and L. Rupp (Eds.) Feminist Frontiers*.

Gaard, G. (2011). Ecofeminism revisited: Rejecting essentialism and re-placing species in a material feminist environmentalism. *Feminist Formations*, *23*(2), 26–53.

Giraldo, I. (2016). Coloniality at work: Decolonial critique and the postfeminist regime. *Feminist Theory*, *17*(2), 157–173. https://doi.org/10.1177/1464700116652835

Gunkel, D. J. (2015). The rights of machines: Caring for robotic care-givers. In *Machine Medical Ethics* (pp. 151–166). Springer.

Gunkel, D. J. (2018a). *Robot rights*. MIT Press.

Gunkel, D. J. (2018b). The other question: Can and should robots have rights? *Ethics and Information Technology*, *20*(2), 87–99.

Haraway, D. (1989). *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. New York: Routledge.

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*. Routledge.

Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.

Hayles, N. K. (2008). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Hertzmann, A. (2019). Aesthetics of Neural Network Art. *ArXiv Preprint ArXiv:1903.05696*.

Irani, L., Vertesi, J., Dourish, P., Philip, K., & Grinter, R. E. (2010). Postcolonial computing: A lens on design and development. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1311–1320. ACM.

Jenkins, R. (2014). *Social identity*. Routledge.

Johnson, B. (2019, April 6). Hey Google, sorry you lost your ethics council, so we made one for you. Retrieved August 11, 2019, from MIT Technology Review website: https://www.technologyreview.com/s/613281/google-cancels-ateac-ai-ethics-council-what-next/

Jones, R. (2015). *Personhood and Social Robotics: A psychological consideration*. Routledge.

Kaplan, D. M. (2009). *Readings in the Philosophy of Technology*. Rowman & Littlefield Publishers.

Katz, E. (2000). Against the inevitability of anthropocentrism. *Beneath the Surface: Critical Essays in the Philosophy of Deep Ecology*, 17–42.

Kera, D., Block, A., & Link, L. (2009). Digital memorials & design for apocalypse: Towards a non-anthropocentric design. *Communications and New Media Programme National University of Singapore Faculty of Arts & Social Sciences*.

Keyes, O., Hoy, J., & Drouhard, M. (2019). Human-Computer Insurrection: Notes on an Anarchist HCI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 339. ACM.

Khader, S. J. (2018). *Decolonizing Universalism: A Transnational Feminist Ethic*. Studies in Feminist Philosophy.

Konopka, A. (2013). Public, Ecological and Normative Goods: The Case of Deepwater Horizon. *Ethics, Policy & Environment*, *16*(2), 188–207.

Koops, B.-J. (2006). *Should ICT Regulation Be Technology-Neutral?* (SSRN Scholarly Paper No. ID 918746). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=918746

Lakoff, G., & Johnson, M. (1980). Metaphors we live by: University of Chicago press. *Chicago, IL*.

Latonero, M. (2018). *Governing Artificial Intelligence: Upholding Human Rights and Dignity* (p. 38). Retrieved from Data&Society website: https://datasociety.net/wp-

content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Latour, B. (2003). Do you believe in reality? News from the trenches of the science wars. *Philosophy of Technology: The Technological Condition, Blackwell Publishing Ltd*, 126–137.

Lawhead, J. (2015). Structural Modeling Error and the System Individuation Problem [Preprint]. Retrieved August 12, 2019, from http://philsci-archive.pitt.edu/11971/

Leopold, A., 1886-1948. (1949). *A Sand County almanac, and Sketches here and there*. Retrieved from https://search.library.wisc.edu/catalog/9910065222502121

Lewis, S. L., & Maslin, M. A. (2015). Defining the Anthropocene. *Nature*, *519*(7542), 171–180. https://doi.org/10.1038/nature14258

Lupinacci, J. (2015). Recognizing Human-Supremacy. *Anarchism and Animal Liberation: Essays on Complementary Elements of Total Liberation*, 179.

Metzinger, T. (2019, August 24). Ethics washing made in Europe. Retrieved July 31, 2019, from Der Tagesspiegel website: https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Midgley, M. (2003). *Utopias, Dolphins and Computers: Problems in Philosophical Plumbing*. Routledge.

Mills, C. W. (2005). "Ideal Theory" as Ideology. *Hypatia*, *20*(3), 165–184. https://doi.org/10.1353/hyp.2005.0107

Mills, C. W. (2011, April 4). The Political Economy of Personhood « On the Human. Retrieved July 29, 2019, from On the Human website: https://nationalhumanitiescenter.org/on-the-human/2011/04/political-economy-of-personhood/

Musiał, M. (2017). Designing (artificial) people to serve – the other side of the coin. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(5), 1087–1097. https://doi.org/10.1080/0952813X.2017.1309691

Neely, E. L. (2014). Machines and the moral community. *Philosophy & Technology*, *27*(1), 97–111.

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

Nocella, A. (2012). Challenging whiteness in the animal advocacy movement. *Journal for Critical Animal Studies*, *10*(1), 142–154.

Nocella, A., White, R. J., & Cudworth, E. (2015). *Anarchism and animal liberation: Essays on complementary elements of total liberation*. McFarland.

Nussbaum, M. C. (1998). *Cultivating humanity*. Harvard University Press.

Nussbaum, M. C. (2010). *Not for profit: Why democracy needs the humanities* (Vol. 2). Princeton University Press Princeton, NJ.

Oliveira, H. G. (2017). O Poeta Artificial 2.0: Increasing meaningfulness in a poetry generation twitter bot. *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, 11–20.

Petersen, S. (2007). The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*(1), 43–54.

Petersen, S. (2012). Designing People to Serve. In P. Lin, G. Bekey, & K. Abney (Eds.), *Robot Ethics. The Ethical and Social Implications of Robotics* (pp. 283–298). Cambridge, MA; London: MIT Press.

Prescott, T. J. (2017). Robots are not just tools. *Connection Science*, *29*(2), 142–149. https://doi.org/10.1080/09540091.2017.1279125

Rainey, S. (2016). Friends, robots, citizens? *ACM SIGCAS Computers and Society*, *45*(3), 225–233. https://doi.org/10.1145/2874239.2874271

Reardon, J. (2009). *Race to the Finish: Identity and Governance in an Age of Genomics*. Princeton University Press.

Reed, C. (2007). Taking sides on technology neutrality. *SCRIPTed*, *4*, 263.

Risse, M. (2018, December). *Human Rights and Artificial Intelligence: The Long (Worrisome?) View*. Presented at the Human Rights, Ethics, and Artificial Intelligence: Challenges for the next 70 Years of the Universal Declaration, The Carr Center for Human Rights Policy. Retrieved from https://youtu.be/YniwuPWhHSo

Roden, D. (2014). *Posthuman life: Philosophy at the edge of the human*. Routledge.

Romero, S. (2018, December 31). Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars. *New York Times*. Retrieved from https://www.nytimes.com/2018/12/31/us/waymo-self-driving-cars-arizona-attacks.html

Said, C. (2019, May 26). Kiwibots win fans at UC Berkeley as they deliver fast food at slow speeds. *San Fransisco Chronicle*. Retrieved from

https://www.sfchronicle.com/business/article/Kiwibots-win-fans-at-UC-Berkeley-as-they-deliver-13895867.php

Said, E. W. (2004). *Humanism and democratic criticism*. Columbia University Press.

Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B., … Dario, P. (2010, October 15). *How safe are service robots in urban environments? Bullying a robot*. 1–7. https://doi.org/10.1109/ROMAN.2010.5654677

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, *39*(1), 89–119.

Smith, D. H., & Zeller, F. (2017). The Death and Lives of hitchBOT: The Design and Implementation of a Hitchhiking Robot. *Leonardo*, *50*(1), 77–78.

Spiel, K., Keyes, O., & Barlas, P. (2019). Patching Gender: Non-binary Utopias in HCI. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*, 1–11. https://doi.org/10.1145/3290607.3310425

Steiner, G. (2010). *Anthropocentrism and its discontents: The moral status of animals in the history of western philosophy*. University of Pittsburgh Pre.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A.-K. (2017). *Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts*. 11.

Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies*, *127*(1), 109–166.

Suárez-Serrato, P., Velázquez Richards, E. I., & Yazdani, M. (2018). Socialbots Supporting Human Rights. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, 290–296. https://doi.org/10.1145/3278721.3278734

Thomas, V., Remy, C., & Bates, O. (2017). The Limits of HCD: Reimagining the Anthropocentricity of ISO 9241-210. *Proceedings of the 2017 Workshop on Computing Within Limits*, 85–92. https://doi.org/10.1145/3080556.3080561

Turing, A. (1947). Lecture to the London Mathematical Society. In B. Carpenter & R. Doran (Eds.), *A. M. Turing's ACE Report of 1946 and Other Papers*. Cambridge, MA: MIT Press.

Valentin, G. (2014). *From HCI to ACI: User-centered and Participatory design in Canine ACI*. Georgia Institute of Technology.

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the Reasons for Making Artificial Moral Agents. *Science and Engineering Ethics*, *25*(3), 719–735. https://doi.org/10.1007/s11948-018-0030-8

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Penn State Press.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Vincent, J. (2017, October 30). Pretending to give a robot citizenship helps no one. Retrieved August 11, 2019, from The Verge website: https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudi-arabia-sophia

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Science Fiction Criticism: An Anthology of Essential Writings*, 352–363.

Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003). CAPTCHA: Using hard AI problems for security. *International Conference on the Theory and Applications of Cryptographic Techniques*, 294–311. Springer.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., … Schwartz, O. (2018). *AI now report 2018*. AI Now Institute at New York University.

Wilde, O. (1891). The Soul of Man Under Socialism. *Fortnightly*, *49*(340), 292–319.

Williams, D. (2019a). Consciousness and Conscious Machines: What's At Stake? *AAAI Spring Symposium: Towards Conscious AI Systems*. Presented at the Stanford CA. Retrieved from http://ceur-ws.org/Vol-2287/paper5.pdf

Williams, D. (2019b). Heavenly Bodies: Why It Matters That Cyborgs Have Always Been About Disability, Mental Health, and Marginalization. *Mental Health, and Marginalization (June 8, 2019)*.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 121–136.

Wise, T. (2005, August 13). Animal Whites: PETA and the Politics of Putting Things in Perspective. Retrieved August 3, 2019, from Tim Wise website: http://www.timwise.org/2005/08/animal-whites-peta-and-the-politics-of-putting-things-in-perspective/

Young, I. M. (1988). Five Faces of Oppression. *Philosophical Forum*, *19*, 270–290. Wiley.