The Journal of Sociotechnical Critique

Volume 1 | Issue 1 Article 5

May 2020

Human supremacy as posthuman risk

Daniel Estrada
New Jersey Institute of Technology, djestrada@gmail.com

Recommended Citation

Estrada, D. (2020). Human supremacy as posthuman risk. *Journal of Sociotechnical Critique*, *1*(1), 1–40. https://doi.org/10.25779/j5ps-dy87

Human supremacy as posthuman risk

Cover Page Footnote

A version of this paper was presented at Computer Ethics—Philosophical Enquiry (CEPE) 2019, at Old Dominion University in Norfolk, Virginia, in the United States of America. It was selected to be included in a special issue devoted to papers from the conference, contingent on successful normal blind peer-review processes. Due to disruptions in article revision processes caused by the COVID-19 pandemic, and the very different ways these disruptions have affected different authors, JStC has chosen to publish articles from this special issue as they are ready, rather than waiting to publish them all together. 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, Kimberly Prince, 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. This paper is dedicated to Felicia Estrada, who was a good person.

Human supremacy as posthuman risk

Daniel Estrada
New Jersey Institute of Technology

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

Keywords: AI ethics, posthuman, animal studies, robot studies, ideal theory

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 proposal describing "An ethical framework for a good AI society" (Floridi et al., 2018), and Stanford's "Institute for Human-Centered AI," whose welcome page proudly proposes that "AI is to serve the collective needs of humanity."[3]

In these contexts, centering the "human" as the explicit focus of normative concern projects the appearance of an inclusive framework of shared values and common interests that guide the collective use of AI. More subtly, these proposals consistently frame AI as categorically subservient to human interests. Unsurprisingly, these proposals don't define the scope or content of the "human," how membership in this category is to be determined, or what underpins its role as a focal norm in AI policy. Still, AI is regularly used to target, manipulate, incarcerate, and exploit vulnerable populations, as seen in the use of facial recognition technologies in

[1] https://www.aiforhumanity.fr

[2] https://www.humane-ai.eu/

[3] https://hai.stanford.edu/

policing and military operations, election tampering and voter manipulation on social media, and software that automates criminal sentencing, loan approval, hiring decisions, and so on (Williams, 2019b; Spiel et al., 2019; Buolamwini & Gebru, 2018; Eubanks, 2018; Noble, 2018; Asaro, 2016; Angwin et al., 2016; Benjamin, 2016). Given these realities, 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 (Jackson, 2020; Deckha, 2012; Mills, 2011; Hayles, 2008; Said, 2004; D. Haraway, 1989, 1991), animal rights and environmental ethics (Belcourt, 2015; Gaard, 2011; Steiner, 2010; Katz, 2000); STS, HCI, and design theory (Thomas et al., 2017; Kera et al., 2009; Latour, 2003), 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 (Khader, 2018; Giraldo, 2016). 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 paper "Robots should be slaves" (Bryson, 2010a). In this and several other essays (Bryson & Kime, 1998, 2011; Bryson, Diamantis, & Grant, 2017; Bryson, 2011, 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 and domestic servitude. 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 recent scholarship on posthumanism and anthropocentrism, and especially through the lens of 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 ethical commitment 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 comment helps locate the matrix of oppression, artificial agency, and group identity addressed in this critical evaluation of human supremacy. 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. (2019, p. 56ff)

This paper is presented in solidarity with those 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 take on 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 helps uncover 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 justice 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." The ideologies of oppression and abject domination indicated by the term "supremacy" 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 (Nocella et al., 2015; Nocella, 2012; Gaard, 2011; Wise, 2005; Armstrong, 2002). 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] Animal studies scholar Anthony Nocella (2012) argues that

[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

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 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). Greta Gaard (2011) points to efforts in ecofeminist thought that foreground the intersections of race, 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). This does not imply that human rights and animal advocacy work are always in alignment. Zakiyyah Iman Jackson notes that,

> animal advocacy projects that seek greater legal protection for the Great Apes and more strenuous criminal prosecution for those who transgress protective laws find themselves at odds with impoverished peoples in African nations that have been burdened by World Bank and IMF policies. (Jackson, 2020, p. 15ff)

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 forms of oppression "if that

> one must possess in order to focus on animal rights as opposed to, say, surviving oneself from institutional oppression. (Wise, 2005, as quoted in Nocella, 2012)

[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 would like to acknowledge that this research was done during the tragic 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. Revisions for this article were completed during a pandemic, and amidst worldwide Black Lives Matter demonstrations in protest of police violence and systematic injustice.

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 well to the potentially insensitive comparisons the term "human supremacy" invites in the context of artificial agents and AI.[6] It is important to keep in mind 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, capitalism, and other forms of systemic oppression (Keyes et al., 2019; Bardzell & Bardzell, 2015; Irani et al., 2010). A critical inquiry into the institutionalized abjection that develops around the use and deployment of robots and AI, or what might be called *critical robot studies*, addresses a particularly salient form of anthropocentric ideology, and aims to resist its careless use in the defense of systemically oppressive practices in AI ethics. 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, neoliberal colonialism, 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). We do not hope to speak on behalf of or "liberate" robots, nor to merely appropriate the language and culture of activist movements. Instead, we seek to contribute to the struggle against all forms of oppression by examining one manifestation of a tactic that impacts humans and nonhumans alike; namely, the political classification of agents by *type*, and the systematic privileging of groups based on essentializing, hierarchical ontologies (Jackson, 2020; 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 ethics in order to name a nefarious mode of classification politics that situates the "human" as the focus of systemic privilege. While humanist or anthropocentric framings can be found throughout the AI ethics literature, our project is not condemn the vocabulary of the "human" wherever it appears; as Jackson says, "To render one's humanity provisional, where the specter of

[6] The term "AI" and "robot" are used here to include all technologies addressed under the labels artificial intelligence, machine learning, and robotics, including autonomous vehicles, drones, and weapons, IoT and "smart" appliances, social media bots and other artificial software agents (anthropomorphic or not), expert systems and efficient database management architectures, and related technologies. The vocabulary and taxonomy for identifying and distinguishing between artificial agents playing various roles remains unsettled.

nullification looms large, is precisely the work that racism does" (Jackson, 2020, p. 16). Rather, our goal is to uncover the ideologies of oppression and abject domination that often informs the appeal to humanism as an ethical ground. Used in this way, the term retains some of its original meaning and use. Nevertheless, the critique of human supremacy in AI presents several unique challenges that distinguish it from the animal advocacy case. One difference is that while anthropocentrism in environmental policy can be subtle and may require critical or interpretive efforts to "recognize" (Lupinacci, 2015), in AI human supremacy is often *overt*, with the "human" presented as an explicit basis for political alliance, as we will see in Bryson's view. To this extent, the term "human supremacy" functions less as an accusation of covert oppressive behavior by analogy to racial oppression, and more as a precise description of an ideology framed in its proponent's own terms.[7]

It is worth considering why human-centered politics is so broadly welcomed in AI ethics, despite the otherwise 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 well-established 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's critique of ideal theory. 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 or critical advocacy. If animal advocacy is primarily motivated by animal suffering, and if robots cannot suffer, this would undermine the possibility that robots could stand in need of the sort of political activism seen in the animal advocacy movement. Such arguments can be resisted on several grounds. Environmentalists since at least 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 warrant a focal role in

[7] The vocabulary of "supremacy" has become popular tech jargon to indicate superiority in some practical domain, as with the so-called race to "quantum supremacy" (Arute et al., 2019). This "supremacist" framing pitting humans against bots can be seen regularly in AI discussions around competitive games, as with bots competing for "poker supremacy" (Gibney, 2017), or bots that challenge "human supremacy at chess" (Müller & Schaeffer, 2018).

our norms and practices (Konopka, 2013). Kate Darling 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 U.S. 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 emphasis on conscious experiences in AI privilege a Western European and predominantly Christian perspective on artifacts and their relationship to nature, ethics and society—a perspective that is not universally shared (Williams, 2019b; Gunkel, 2018a; Jones, 2015). Assuming this perspective marginalizes non-Western traditions that do not operate on an anthropocentric hierarchy of value, such as found in Shinto (Geraci, 2006), African and African diasporic (Jackson, 2020; Metz, 2017; Horsthemke, 2017), and First Nations (Pierotti & Wildcat, 2000) philosophy. Finally, there are the ongoing discriminatory practices in which biological factors are treated as scientific justification for institutional oppression based on race, gender identity, sexual orientation, disability, or other essentializing characteristics (Jackson, 2020; Appiah, 2018). Taken together, these considerations suggest that biological factors alone 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 internal states sufficiently "like ours" to warrant social status and legal recognition (Wittkower, forthcoming; Danaher, 2019; Williams, 2019a, 2018; Gunkel, 2018b; Schwitzgebel & Garza, 2015).[8] The hypercritical focus on the machine's experiences (or lack thereof) points to another important distinction between animals and AI: artificial agents already participate in a variety of sociopolitical contexts that were formerly the exclusive domain of humans. There already exist vibrant online communities building, publishing, and critically assessing the work of bots that write poetry and fiction, create digital images and videos, compose music, and produce other forms of 'artistic' expression (Hertzmann, 2019; Oliveira, 2017; Compton et al., 2017; Gilani et al., 2017). Perhaps most well-known are the artificial "influencers" like Lil Miquela who model fashion brands to millions of followers on social media

[8] These concerns predate Turing's (1950) proposed "imitation game". Turing attempts to respond to these concerns by directing 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 more popular reading of Turing's test runs afoul of Nocella's recommendation to avoid comparing experiences. This is not to suggest that Turing's test is morally wrong, but to notice that it creates social circumstances that are especially hostile to the possibility of recognizing and respecting artificial agents (Hayes & Ford, 1995).

(Blanton & Carbajal, 2019). These digital communities highlight the already overlapping sociopolitical circumstances of human and artificial agents, which are not predicated on some shared biological or evolutionary 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 for collaboration provides resources to resist biocentrism, anthropocentrism, and other essentializing hierarchies as they appear in the discourse around robots, AI, and artificial agency, despite important differences between human and artificial agents.

Moreover, unlike nonhuman animals, it is reasonable to expect the capacities of artificial agents will continue to advance rapidly on relatively short timescales, at least within some domains. The possibility for radical near-term changes in the agential capacities of robots suggests their sociopolitical status will likewise remain unsettled. Peter Asaro notes, "At some point in the future, robots might simply demand their rights" (Asaro, 2006, p. 12). However, while articulating these demands may require fundamental changes in the capacities of artificial agents, it will also require an ethical and interpretive change in our capacities as a society to respect such demands as potentially legitimate political acts. These changes in the social imagination don't require that we wait for technologies "worthy" of basic consideration. Alan Turing articulated this perspective in his plea for "fair play for machines" (Turing, 1947; Estrada, 2018a) just as the first general purpose computers inspired by his work were being constructed in research labs. 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 AI ethics

Joanna Bryson's contributions to the AI ethics literature, especially the article "Robots should be slaves" (Bryson, 2010; henceforth RSBS), provides a useful case study for the critical examination of human supremacist ideology in AI ethics. RSBS is notable (though, as discussed earlier, hardly unique) for its unfortunate invocation of the historical institution of slavery as a model for AI ethics, where "robots should be servants you own" (p. 3). Bryson would later apologize for this framing (Bryson, 2015). Still, the set of considerations, drawn mostly from evolutionary psychology, that lead to proposing slavery as a model for AI ethics continues to inform Bryson's work not only as a scholar but also as

a high-level policy consultant for both government and industry. This section will present a critical reading of RSBS and related work, to shed light on the motivation and perspective they develop and the influence they have had on AI ethics discourse and industry policy. To be clear, Bryson's work is not singled out for being egregiously problematic or offensive. On the contrary, Bryson clearly articulates a mainstream liberal humanism that is widely endorsed in the field, for which she has received broad support and recognition. Given the prominence of these views, it is important to take seriously the ways in which they rely on or perhaps even require the logic and essentializing hierarchies of white supremacy. In this section we detail Bryson's straightforward embrace of the logic of institutionalized slavery in RSBS and related papers to address these deeper themes in the AI ethics discourse.

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. An edited version of this paper is published in 2011 under the title "Just an artifact: Why machines are perceived as moral agents" (Bryson & Kime, 2011), suggesting some continuity of views over this time, the period during which RSBS was written, presented, and published. As such, these papers give insight into Bryson's early views on AI ethics independent of the overt analogy with slavery developed in RSBS. Bryson & Kime's motivation in these papers is to address certain "exaggerated fears" (Bryson & Kime, 1998, p. 385) from Vernor Vinge and other early proponents of the Singularity hypothesis, who predict that computers might soon surpass human intelligence and take over the world (Vinge, 1993). Bryson & Kime argue that these misplaced fears arise from an "over-identification with machines," a mistake they say is "symptomatic of a larger problem—a general confusion about the nature of humanity and the role of ethics in society" (p. 385). 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. 386). They claim 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" (p. 386, emphasis in original). This feeling of empathy in turn creates a sense of *identification* with the target of our concern. 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" (p. 386). Bryson & Kime infer from this general picture suggests that "self-interest is the root of our ethics" (p. 387).

In the case of the Singularity theorists, Bryson & Kime argue that the mechanisms of social identification have 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. “To form a human society, one needs to value the lives of humans in the community over the lives of other animals” (p. 386). They argue that a mistaken 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” (p. 387*).* They identify two dangers of over-identification: “we may believe the machine to be a participant in our society, which might seriously confuse our understanding of them,” and “we may over-value the machine when making our own ethical judgments and balancing our own obligations” (p. 387). They claim these dangers are not unique to AI, and that all artifacts have “the potential for misuse, either through carelessness or malevolence, by the people who control them.” (p. 385) They dismiss any view that values AI, “to the exclusion of our own existence” as “nihilism” (p. 390). The most substantial citation offered to support these claims is Lakoff & Johnson's *Metaphors we live by* (1980).

There are many reasons for finding these views unsatisfying as an ethical framework. The direct line drawn between evolutionary psychology and ethical obligation is underdeveloped and theoretically implausible (Street, 2006). The relationship proposed between empathy and social identification is, at best, oversimplified (Jenkins, 2014). The view that artifacts are dangerous only through their 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 the high tech industry (Reed, 2007; Koops, 2006; Winner, 1980). However, it is not our goal to present a full scholarly critique of a conference paper from twenty years ago. Instead, our goal is to trace the development of a view that results in the explicit endorsement of slavery as an ethical framework for managing robots. For our purposes, the most relevant features of Bryson’s position in these early papers are the claims that *identification drives moral obligation*, 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. (p. 6)

Bryson's later work emphasizes that moral systems should be "coherent" (Bryson, 2018a, p. 202), but this would be difficult to achieve if morality were grounded on empathetic identification alone. Given the complex psychological and political realities involved in the production of social identity, it seems unlikely that "over-identification" (with machines or anything else) is a serious threat to the social order. Nevertheless, Bryson & Kime (1998) take the cultivation of appropriate identification practices towards machines be a central concern in AI ethics, one which the proposal in RSBS and later works aim to address.

For the sake of argument, 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 *prima facie* 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. 389). So why not take the supposed identification with machines as evidence of evolving moral obligations? The authors defend their resistance to adopting new ethical standards for machines with two responses, one they describe as "technical," the other "ethical." Their technical response argues that people tend to overestimate the capacities of machines. Their ethical response argues that we already face a resource allocation problem with other artifacts like "fine art and political institutions," (p. 387) where investments might draw resources away from more pressing human interests.

Neither response addresses the issue at stake, which is how to decide which identification practices (and which identities) are appropriate and which are mistaken. 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 not. 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 failures. The ethics of identification simply are not settled by the accuracy of the predictions those identities generate.

This technical response is confusing, given that their proposal contains 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 raise 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? (Wittkower 2020; Carpenter, 2016; Darling, 2015) By insisting on a principled basis that the identification with machines is a *conceptual mistake*, Bryson & Kime (1998) cut off these possibilities and effectively limit the ethical discourse in AI to controlling the frequency and impact of these "anthropomorphic fallacies" (p. 390). In this spirit they claim, "The issue of forming identity is now more than ever an issue for public education," (p. 391) implying a need for institutionalized policies to control how social identities are formed and who we identify with. This interest in controlling the identification practices and empathetic responses people might form towards machines is a central pillar of the argument developed in RSBS.

Bryson & Kime's (1998) second "ethical" response reveals an important assumption in Bryson's ethical perspective: that the evolutionary dynamics of obligation are often *zero sum*, and that developing new obligations towards AI would entail fewer obligations to and empathy with humans, animals, and society generally. The risk is not simply that we mistakenly 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 social cohesion depends on our empathy and obligations towards other humans, the over-identification with AI is not merely inappropriate or distasteful; it is a direct threat to the social order. The authors emphasize that this threat is not unique to AI, pointing 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 fine art threatens social cohesion (and ultimately our evolutionary success!) by potentially generating more empathy and resources 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 fine 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, and empathetic society because of the resources we invest in public art. But for Bryson & Kime, producing fine art is a social liability, a cost we accept, like car accidents, because of the pleasure and convenience those artifacts bring us. This zero-sum perspective on obligation is another major pillar of Bryson's reasoning in RSBS and later essays.

"Robots should be slaves" (Bryson, 2010a) was 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). Together with the revised "Just an artifact" (Bryson & Kime, 2011), these papers all expand on the themes developed in "Just Another Artifact" (1998), emphasizing the ethical risks posed by the over-identification with AI. Bryson's work in this period is not aimed at the subjugation of robots as such. Instead, her work tries to correct what she sees as the conceptual confusion generated by an inappropriate identification with machines, a mistake propagated by science fiction narratives and self-described "futurists" whose doomsday scenarios had become popular in tech journalism around this time. Bryson recommends against the use of anthropomorphic robots in industrial settings, and for policies that establish an unambiguous ethical hierarchy that situates responsibility and ethical priority with humans over machines. In 2010, Bryson participated in a joint EPSRC/AHRC (Engineering and Physical Sciences Research Council/Arts and Humanities Research Council) retreat with industry leaders and policy experts in the UK. This retreat produced a set of "Principles of Robotics" (Boden et al., 2017), which reinforces Bryson's themes concerning the identification with machines. Of the five rules laid out in the document, the first four largely concern *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)

The principles propose industrial policies that clearly distinguish between the capacities of robots and humans, advocating a purely instrumental perspective on the former, and exclusively restricting the discussion of agency and responsibility to the latter. As a policy document, these principles give the misleading impression that the primary ethical risks presented by industrial applications of AI and robotics are ontological confusions over their agential status.

The rationale driving RSBS can now be brought into focus. Bryson's proposal takes for granted not only that humans should not be treated as slaves, but also that *no one would identify or empathize with slaves*. For Bryson, the "slave" is quintessentially an inhuman "other," an archetype that is characteristically beyond the scope of moral consideration or empathetic identification. By calling *robots* "slaves," and by treating the categorical distinctions between humans and robots with the same moral weight given to the distinction between humans and slaves, Bryson hopes to counteract the excessive empathy we might feel with robots through the attitude of abject disregard we "should" feel towards slaves. The call for

robot slavery directly answers Bryson's concern across several papers to establish rigid identity hierarchies that formalize the ethical non-status of machines. In RSBS, Bryson repeats earlier claims that "our identity confusion results in somewhat arbitrary assignments of empathy" (Bryson, 2010a, p. 4), and lists a set of costs for both individuals and institutions associated with the over-identification with AI (p. 5). Bryson describes "being too generous with personhood" as a "moral hazard" (p. 7). She rehearses the zero-sum reasoning,[9] 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 empathetic identification with machines, or how to design technologies that maximize these benefits.

Instead, Bryson uses the 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

[9] Bryson et al. (2017) recognizes the potential problems with zero-sum ethical 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. (p. 275)

Notice that this argument conflates the potential harms that result from the creation of new legal individuals with the potential harms that may arise from expanding the category of personhood.

employees were essential" (p. 8).[10] 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).

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 cause 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. Since robots are not moral agents, 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 extended mind theory encourages us to identify strongly with our machines (Ahuvia, 2005), or how this might be inconsistent with her proposal for robots as slaves.

It should go without saying that the appeal to institutionalized slavery and servitude as "good and useful, … right and natural" is profoundly insensitive and simply in poor taste. It also highlights a deep theoretical failure in Bryson's ethics. Just as with the *Mechanix Illustrated* comic from 1965 quoted earlier by Ruha Benjamin, 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 to the substantial challenges involved in grounding ethical policy in the collective construction of social identity. Although Bryson makes token gestures to recognize the historical cruelty of racialized slavery, she does not consider how the metaphor of slavery might be interpreted by those who identify more with slaves rather than with slaveholders. She also fails to consider how a defense of slavery as a political institution might present a greater hazard to the moral imagination than an overidentification with robots. Fundamentally, Bryson does not think the problem with slavery is the ideology of domination and oppression it represents; the problem with slavery is that we were enslaving the wrong things.

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

RSBS has received substantial scholarly attention in the AI ethics literature (Agar, 2019; van Wynsberghe & Robbins, 2019; Gunkel, 2018b, 2015; Frank & Nyholm, 2017; Musiał, 2017; Prescott, 2017; Rainey, 2016; Coeckelbergh, 2015; Neely, 2014). While much of this literature is critical of Bryson's insensitive language, few engage her theoretical approach from the perspective of standpoint epistemology or critical race theory to reflect on the role that racialized hierarchies play in the AI ethics discourse. Consequently, the human supremacist framing developed in RSBS continues to find endorsement in the literature. For instance, Birhane & van Dijk recently state their "full agreement" with Bryson's focus on human well-being in RSBS, and disagree only to the extent that "one cannot dehumanize something that wasn't human to begin with" (Birhane & van Dijk, 2020). The primary lesson Bryson has drawn from this feedback is that "you cannot use the term 'slave' without invoking its human history" (Bryson, 2015). Bryson has apologized for her insensitive use of the word "slave," but not for the oppressive ideology that language articulates, or the abject disregard it shows for those who do not share her identities or perspectives.

Indeed, Bryson's recent work continues to develop the central perspective of RSBS. While the explicit vocabulary of slavery has been dropped, the focus on institutionalized identification policies and anthropocentric ontologies remains. In "Of, for, and by the people: The legal lacuna of synthetic persons" (Bryson et al., 2017), Bryson claims the "the basic purposes of human legal systems" includes a principle that, "[s]hould equally weighty moral rights of two types of entity conflict, legal systems should give preference to the moral rights held by human beings" (p. 283). Note 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 the character of Bryson's position, she describes it as, "an uncontroversially light thumb on the scale in favor of human interests. Yes, this is speciesism" (p. 283). Since humans and machines are not distinct species, the term 'speciesism' is a misnomer; we suggest the term "human supremacy" as a more appropriate characterization of this position. In "No one should trust AI" (Bryson, 2018b), Bryson argues that trust is a relationship between peers, and since we aren't peers with AI, "no one actually *can* trust AI" (para. 2, emphasis in original). In "Patiency is not a virtue: The design of intelligent systems and systems of ethics" (Bryson, 2018a), Bryson argues that "where possible there should be minimal restructuring of existing norms" (p. 17), and that making robots deserving of moral consideration "could in itself be construed as an immoral action" (p. 16). These papers consistently argue that human social cohesion is an evolutionary imperative that must be met with institutionalized hierarchies which systemically privilege the "human." While the outright appeal to slavery has been suppressed, the logic of human supremacy has, if anything, become more pronounced.

In these papers, Bryson's concern shifts from the overidentification with machines as such to the more indirect threat that malicious corporate actors could anthropomorphize machines to exploit both our empathetic biases and legal loopholes around personhood. While the ethics of corporate personhood is a well-placed concern, Bryson conflates this with the supposed challenges of anthropomorphism and over-identification, saying,

> For example, customers could be fooled into wasting resources needed by their children or parents on a robot, or citizens could be fooled into blaming a robot rather than a politician for unnecessary fatalities in warfare. A corporation could displace responsibility for its decision to use automation rather than human employment onto the automation itself, creating a legal lacuna—a set of far poorer, purely-synthetic entities set up to be held responsible for tax and legal liability (p. 23).

Outside of certain science fiction scenarios, it is not clear how these risks are related. Corporate personhood and liability law do not depend on the psychology of empathetic identification. Bryson does not point to any case where anthropomorphic robots have been used to evade legal liability or establish legal personhood. She does not explain how anthropocentric ontologies or restrictions on anthropomorphic robots would prevent corporate abuses of legal personhood. She takes for granted that human supremacist policies would protect all and only the "correct" moral agents.

Bryson's work in AI ethics extends beyond academic scholarship and coincides with her rise to prominence as a public speaker and high-level policy consultant. In 2017, Bryson was quoted calling the popular humanoid robot Sophia's award of Saudi citizenship "bullshit" (Vincent, 2017). The comment earned a public debate with Sophia creator David Hanson at CogX 2018 (Estrada, 2018a).[11] In 2019, Bryson was selected to sit on the Advanced Technology External Advisory Council at Google (Johnson & Lichfield, 2019). This council drew controversy for the inclusion of Kay Coles James, president of the Heritage Foundation, a conservative think tank known for promoting regressive policies on immigration and LGBTQ rights (Knight, 2019). After an open letter from immigrant and LGBTQ tech workers condemning the council led to Google's termination of the program, Bryson continued to defend the council and her participation in it (Bryson, 2019b). Bryson also defended James's participation on the council, comparing the criticisms she and the council had received to "bullying and shunning" (Bryson, 2019a).

[11] The debate was initially marketed as between Bryson and Sophia the robot, until Bryson objected on social media, and Hanson agreed to take the robot's place on stage. The debate itself largely concerned the virtues and risks of anthropomorphism in robotics. See Estrada (2018a).

Bryson has made clear that she does not share James's political views. We do not raise this minor controversy here to smear Bryson by an indirect association with James's foundation, but instead to reflect on how corporate tech giants like Google cater to entrenched political interests over the interests of people systemically marginalized by those politics. In this context, it is notable that Bryson's public response to the controversy was to defend her work with James and Google over public outcry from the very communities they target. Given Bryson's scholarly concern for the careful construction of social identities as a stalwart against corporate abuses, and the emphasis she gives to human interests as a unifying norm, one might be surprised to find her scolding public criticism in defense of corporations and the xenophobic bigots they woo. However, a close reading of Bryson's work reveals a methodological interest she shares with both Google and the Heritage Foundation: the systematic identification and classification of agents into essentializing hierarchies for purposes of political control. The same white supremacist reasoning that motivates the Heritage Foundation to target immigrants and LGBTQ people as political scapegoats, and which Google and other corporations deploy for targeted advertising, surveillance, and policing (Cave, 2020), Bryson has adapted as an organizing framework for policy and ethics across the field of AI and robotics.

As surveyed in the introduction, Bryson is not alone or unique in her explicit embrace of human supremacy as a moral framework. To some extent, Bryson was in the right place when the AI boom hit to find success in an industry and regulatory climate that was particularly receptive to the vision of human supremacy developed in her work. To address this broader milieu in which Bryson's work finds success, we turn next to the work of Rosi Braidotti and Charles Mills.

Reactionary posthumanism as ideal theory

Bryson argues that humanity is in the grips of an identity crisis. If so, Braidotti's framework of "the posthuman" (Braidotti, 2013) may help diagnose the problem. For Braidotti, posthumanism marks our 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 found in critical theory (Braidotti, 1994, 2002; Foucault, 1977); second, an *analytic* posthumanism that develops through 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). Analytic posthumanism would include, but is not limited to, the transhumanist perspectives that motivate Singularity theory in its many guises: as a science fiction plot device, as a bioengineering design approach, and as a conceptual posit in academic scholarship on future risks. In this sense, analytic posthumanism is an implicit target of Bryson's criticisms of Singularity theory. However, a full critique of analytic posthumanism is beyond the scope of this paper. For our treatment of Bryson's constructive views, this section will focus on Braidotti's discussion of reactionary posthumanism.

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 a product of the socioeconomic condition of globalization, which threatens humanity through the reactionary "plagues" (p. 39) 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" (p. 39*)*. Nussbaum acknowledges the problematic historical use of humanist ideals as a discriminatory or exclusive practice, and responds with a call for a neo-humanism that centers the subjectivity of experience. While Braidotti praises this nod to feminist critiques and methods, she argues that Nussbaum, "reattaches [subjectivity] to a universalistic belief in individualism, fixed identities, steady locations and moral ties that bind" (p. 39*).* Braidotti says that due to 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" (p. 39).

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 is not disagreeing with Nussbaum about the value of a liberal education as such, nor is she asserting the reactionary counter-ideal that liberal humanism is necessarily bad. Rather, her point is to recognize how the humanist 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 corporations, this muddies the narrative of the liberal ideals that the university supposedly represents. Similarly, if the rhetoric of universalist humanism is used to protect narrow and exclusive practices, it undermines the apparent universal appeal of those ideals. In this way, Braidotti argues that Nussbaum's nostalgia for

humanist ideals operates as a defense of the very practices that subvert them.

Charles Mills's (2005) critique of the ideology of "ideal theory" provides tools for thinking through this potentially confusing discursive situation. For Mills, "ideal theory" describes not just the use of idealizations, which to some extent cannot be avoided in theoretical discourse. Instead, ideal theory describes the tendency to rely 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 from an idealized "state of nature," rather than addressing the actual social circumstances in which its institutions operate. Mills claims that ideal theory will typically employ assumptions that idealize human capacities, social institutions, and social ontologies 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 continues,

> 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's terms, Braidotti argues that reactionary posthumanism embraces humanist ideals 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 views. What matters for our purposes is Braidotti's analysis of how a reactionary embrace of humanist ideals in a posthuman context can operate on the ideology of ideal theory. The institutional realities of a corporatized university system have real consequences for the value of higher education, and this influence cannot be elided by appeal to the merits of 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 reliance on ideal theory effectively guarantees that those ideals will never be achieved.

Together, these analyses from Braidotti and Mills help to articulate the critical failures in Bryson's ethics beyond the mere insensitivity of her language. Like Nussbaum's, Bryson's work can be read as an expression of reactionary posthumanism, responding to the contemporary crisis of humanism by asserting nostalgic, elitist ideals for traditional social institutions and ontologies, such as "servants are useful and good." These ideals are themselves justified by idealized models of agency and social cognition, and they aim to minimize disruption of the existing social order, thereby protecting the systems of power that benefit from that order. Where Nussbaum sees ideals of cosmopolitan discourse in a liberal education, Bryson sees ideals of stability and clarity of identity in

institutionalized slavery and strict social hierarchies. Bryson presents herself as speaking on behalf of humanity's interest broadly construed, when in fact her proposals showcase a narrow and privileged perspective that alienates those who don't already share it. In so doing, she neglects the political and psychological realities within which humans and nonhumans share existence.

These analyses also help clarify an interpretive issue in the present critique of human supremacy. Again, it is not our intention to condemn anthropocentrism in AI ethics based solely on an abstract analogy with white supremacy. The problem with human supremacy is not simply that it replicates a superficial hierarchy of structural oppression. Instead, the problem with human supremacy is how it operates as a defense of entrenched power and oppressive ideologies, in a context replete with examples of the systemic abuses of that power and the failures of those ideals. The ontological distinctions proposed between humans and machines are not simply modeled on traditional hierarchies of race and gender, they function to extend, legitimize, and calcify them within AI policy and legal theory. This is the fundamental risk of human supremacist ideology, and it should be recognizable as a risk no matter how one feels about anthropocentrist rhetoric or the moral agency of nonhumans. Identifying robots as the "correct" targets of systemic abjection not only justifies the logic, political utility, and technical infrastructure of oppression. It also demands methods for identifying and classifying humans within that framework, if only so they are not mistaken for robots. This is not unique to Bryson's specific articulation of the view but is endemic to the very logic of human supremacy as a policy solution in AI ethics. Human supremacy attempts to protect humans from systemic abuse by constructing a political framework around systemic abuse and then declaring humans to be a privileged category exempt from that abuse. In this way, human supremacist advocates will inevitably find themselves committed to the project of fitting humans within frameworks of systemic abuse. Likewise, the entrenched powers which benefit from systemic abuse will naturally find the human supremacist defense of stable power hierarchies convenient, even flattering, and will inevitably find themselves promoting its application as ideology in AI ethics.

Nonideal AI ethics

Are there alternatives 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 says, "One defect of ideal theories… is their tendency to redirect our evaluative gazes to the wrong normative phenomena" (Khader, 2018, p. 36). By directing our gaze towards the actual conditions of injustice,

Khader argues that we will be in a better position to address those injustices. Fazelpour & Lipton have recently introduced a nonideal perspective to the AI ethics discourse on algorithmic fairness, arguing that "non-ideal theorizing about the demands of justice is a fact-sensitive exercise" (Fazelpour & Lipton, 2020, p. 10). While this is a step in the right direction, we think the value of a nonideal perspective extends beyond an emphasis on empirical methods. To push the analysis, this section looks to Khader's nonideal universalism for other lessons that might be adapted to an AI ethics context.

Khader's nonideal universalism is developed in the context of a diverse transnational feminist movement, partly to address what she calls "missionary feminism," the idea that feminism requires the universal adoption of Western liberal humanist values (p. 3). Since these values are not universally shared, some critics argue that missionary feminism continues the imperialist project of imposing Western values on colonized communities. On the other hand, rejecting universal values seems to entail a kind of cultural relativism that threatens to undermine feminism as an inclusive normative project. Khader argues that this is a false dilemma; feminists do not have to choose between imperialism and relativism. For Khader, feminism is defined negatively, as an opposition to sexist oppression. A nonideal feminism starts in conditions of injustice and seeks to reduce or eliminate that injustice where it exists. This does not require a shared commitment to Western ideals or any other background of parochial values to serve as a normative ground for feminist solidarity. Khader's approach "rejects the notion that there is a single feminism-compatible cultural form, thereby undermining the idea that an idealized Western culture is the feminist solution" (p. 7). However, this pluralist approach does not commit feminism to a nihilistic cultural relativism. Rather than focusing on disagreements between community ideals considered abstractly, Khader's nonideal approach recognizes that "the effectiveness of strategies for change varies based on the material conditions and moral vernacular(s) of a given context" (p. 4). Thus, a focus on the actual, unjust conditions of sexist oppression opens space for building a consensus toward practical steps that reduce or eliminate sexist oppression, without demanding univocity in the background norms informing this work.

AI ethicists who are sympathetic to the critique of human supremacy developed in this paper may find themselves facing a dilemma similar to the one Khader describes in the transnational feminist discourse. Like missionary feminism, human supremacy continues and expands a Western imperialist ideology that some ethicists will be hesitant to endorse. However, abandoning humanist ideals may seem to leave us in a moral vacuum, without conceptual guidance for protecting the interests of persons and communities under genuine threat of systemic oppression. This concern might simply reflect a parochial Western bias; as mentioned

earlier, many indigenous or non-Western perspectives operate on non-anthropocentric ethical frameworks. Nevertheless, as a practical matter, the international law and policy framework of human rights is not just abstract, ideal theory; it is also an established institutional reality, one which can mobilize material resources and direct international cooperation in the service of real human interests. This is not to say the framework of human rights achieves every ideal it aspires to, but simply that it is an important and useful tool in the toolbox, one which would work better if it were applied more consistently and inclusively, rather than if it were whittled down or eliminated, at least absent better options on the table. If the critique of human supremacy developed in this paper implies that the framework of human rights is an imperialist project, some AI ethicists may be tempted to bite the bullet and admit their preference for an imperialist ethical regime.

However, Khader's discussion of nonideal universalism suggests a way for AI ethicists to avoid the apparent compromise of principle, where we can productively engage an institutional framework like human rights while rejecting the imperialist framings of reactionary posthumanism and human supremacy. Following Khader's nonideal universalism, we might construct a nonideal AI ethics that begins in conditions of systemic oppression and seeks to reduce or eliminate that oppression. This project does not require a monolithic or exclusive idealization of human agency or cognitive capacity. We can use the framework of human rights to protect the interests of actual human communities without treating that framework as a chauvinist characterization of the ideal rights-bearing agent, one which might justify the exclusion, abjection, or oppression of other humans or nonhumans. A nonideal AI ethics rejects the notion of a single ethics-compatible cultural form of agency or capacity, thereby undermining the idea that an idealized Western anthropocentrism is the ethical solution.

This opens space for rethinking the sociotechnical matrix of human and artificial agents in dynamic political terms, without being misdirected by an impulse to comparison, exclusion, or hierarchy. Robots are situated within existing sociopolitical structures to materially extend and reinforce systems of domination and control, but they also encounter these systems as a practical constraint on their operation, and so can also be mobilized to resist, dismantle, and repurpose these systems. As such, robots may already figure within normative communities as having varying degrees of agency, complex social alliances, and relationships with other human and nonhuman agents. A critical discourse on robots and AI in nonideal conditions can recognize how overlapping structures of institutional oppression situate robots as both agents *and* targets of power—as agents whose identity and operation must be made available for inspection, public scrutiny, and abuse (Romero, 2018; Smith & Zeller, 2017; Brscić et al., 2015; Salvini et al., 2010). To some extent, the actual circumstances of these arrangements and the moral vernaculars of the communities

involved should play a role in our collective assessment of the robot's impact and value.

There are already efforts to reimagine the complex interdependent social and political relationships between human and nonhuman agents (Rahwan et al., 2019; Rainey, 2016). Crawford & Joler break down the manufacturing, labor, and supply chains involved in the production and distribution of Amazon's Alexa, describing, "a vast planetary network, fueled by the extraction of non-renewable materials, labor, and data" (Crawford & Joler, 2018). They continue,

> A full accounting for these costs is almost impossible, but it is increasingly important that we grasp the scale and scope if we are to understand and govern the technical infrastructures that thread through our lives. (para. 5)

The human costs of this technical infrastructure are visible in well-known cases like the use of low-paid crowdsourced labor in machine learning (Lung, 2012). They are also visible in more recent variations, such as Kiwibots, a robotics company building delivery service robots in the Bay Area. Kiwibots farms out the robot's control task out to human operators in Colombia who are paid less than $2 an hour, which owners claim is more than the local minimum wage (Said, 2019). Such cases complicate the idealized moral ontologies constructed around human individuals and robot abjection. As Jackson (2020) says, "The more "the human" declares itself "universal," the more it imposes itself and attempts to crowd out correspondence across the fabric of being and competing conceptions of being" (p. 32). Resisting this imposition requires more than rejecting the vocabulary of the human, it requires the moral courage to imagine alternative relationships with being.

We end the paper with two brief examples involving the use of bots by activist communities (Savage et al., 2016). The term "bot" typically has a negative connotation in social media spaces, associated with spam, trolling, and other malicious uses. This animosity has led to the word "bot" being used as a slur or insult to attack the credibility of other people online (Roth, 2018). But some bots are neutral or even helpful, such as bots that automatically report earthquakes or tell jokes. Scholars have attempted to systematically classify robot kinds into benign or malicious varieties (Stieglitz et al., 2017). However, this idealized project runs into immediate challenges in any practical setting. So-called "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, p. 1). Suárez-Serrato et al. (2018) discuss the use of bots by human rights activists in Mexico organizing around the #Tanhuato hashtag to evade state censorship. They write,

> It is important to pause here and notice that in an instance like this it is not a clear matter whether these bots were benevolent or malignant. It is a matter of perspective. From the point of view of the Mexican armed forces, these bots are acting against their honor. From the point of view of the CNDH they are promoting access to a report of human rights abuse. (p. 2)

The #Tanhuato activists illustrate how the classification and alignment of bots is already a political issue with direct implications for human rights. This discussion does not hinge on armchair musings about far future technologies, or misconceptions over machine agency. The discussion arises from the actual conditions of injustice faced by persons under threat of a surveillance state, and the role bots can play in navigating these conditions. The crude algorithmic agency of Twitter bots evades state censorship and retaliation in ways that contribute to resistance efforts; the fate of these bots might even serve as synecdoche for the conditions of injustice themselves, where silencing bots *is* silencing people.

The #botALLY community provides another example where the social status of bots is an explicit object of political concern. When an update to Twitter's automation policy threatened to remove many prosocial bots from the network, a community of bot developers successfully organized around the #botALLY hashtag to pressure the company to change its policy and allow certain bots to operate (smith, 2017). The developers stressed not only the positive role these bots play in the community, but also the bot developer's responsibility for the bot's operation and the culture they produce. One developer and organizer, Darius Kazemi, built a bot that tweets mashups composed by swapping the subjects from two different headlines[12]. Occasionally, the mashup involved subjects of different genders, resulting in automated headlines that appeared to be making transphobic insults. On receiving feedback from the community, Kazemi designed a word filter that would check for certain slurs or insults before tweeting the results (Kazemi, 2015). He explained the importance of an ethical code to the #botALLY community, saying, "I just don't want my bots doing things that I wouldn't do myself" (smith, 2017, para. 6).

The #botALLY community shows how a close identification with robots can motivate developers to take greater responsibility for their bots and the cultures they (re)produce. This follows from an ethic that recognizes robots as operating within a community for which its members take responsibility. This recognition does not come at the expense of human interests; on the contrary, identifying with machines puts the community in a better position to address the interests of all its members. Like the #Tanhuato activists, this recognition does not trade on anthropomorphic exaggerations of machine agency. In both cases, the use of bots demonstrates a technically sophisticated appreciation of the agential relationships between humans, bots, and the systems of power they

[12] https://twitter.com/TwoHeadlines

operate within. Kazemi says #botALLY has "always been a bit tongue in cheek, not so much as friends of robots qua robots but rather as a banner for a kind of white-hat art-bot maker" (personal communication, 2019 April 25). This playfulness with the boundaries of agency helps open the space for critical reflections on a developer's responsibility for their bots.

To these ends, Kazemi and the #botALLY community have produced guidelines for bot developers looking to make prosocial bots (O'Leary 2016, Kazemi 2013). These guidelines include principles like "A bot is an extension of its creator's will" and "Bots should punch up" (O'Leary, 2016; Richardson, 2013). Similar principles are elaborated in Microsoft's guidelines for "Responsible bots" (Cheng, 2018), which includes suggestions like "Design your bot so that it respects relevant cultural norms and guards against misuse" (p. 2), "Ensure your bot treats people fairly" (p. 3), and "Ensure a seamless hand-off to a human where the human-bot exchange leads to interactions that exceed the bot's competence" (p. 2). Microsoft's guidelines come two years after their public failure developing the chatbot Tay.ai, which quickly learned to repeat racist hate speech on social media (Neff, 2016). Commenting on the Tay.ai controversy, Schlesinger et al. (2018) echoes Haraway's call to "stay with the trouble", arguing that

> Critical issues cannot be addressed through neat separations between what people do and how machines operate. In determining where we go from here, we have to hold onto the complexities of our lived experiences, refusing to reduce the world into something that is uniform or singular. (p. 9)

They go on to recommend ways of mitigating the harms caused by chatbots, including developing "bots that are capable of recognizing and responding to race talk in the near future" (p. 9), which they admit "is no small task and there is no silver bullet" (p. 9).

These recommendations around the development and use of prosocial robots can be understood as applications of a nonideal AI ethics, focused on mitigating real short-term harms while avoiding idealized, monolithic, or parochial solutions. The principles are generous with machine agency and sensitive to fluid exchanges between humans and machines, while taking seriously the developer's responsibility for their robots and the impact they have on the social dynamics in which they are embedded. Far from utopic, these principles start from a recognition of existing conditions of injustice and systemic oppression, and they seek to build robots that are responsive to these conditions and operate in ways that minimize or eliminate those injustices. For those of us committed to collective empowerment without domination, these recommendations point us to ways of including robots in this project.

References

Agar, N. (2019). How to treat machines that might have minds. *Philosophy & Technology, 33,* 269-282. 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, May 23). Machine bias. *ProPublica,* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J. C., Barends, R., Biswas, R., Boixo, S., Brandao, F. G. S. L., Buell, D. A., Burkett, B., Chen, Y., Chen, Z., Chiaro, B., Collins, R., Courtney, W., Dunsworth, A., Farhi, E., Foxen, B., … Martinis, J. M. (2019). Quantum supremacy using a programmable superconducting processor. *Nature*, *574*(7779), 505–510. https://doi.org/10.1038/s41586-019-1666-5

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*. Morgan & Claypool. https://doi.org/10.2200/S00664ED1V01Y201508HCI031

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*. Wiley.

Birhane, A., & van Dijk, J. (2020). Robot rights? Let's talk about human welfare instead. *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 207–213).

Blanton, R., & Carbajal, D. (2019). Not a girl, not yet a woman: A critical case study on social media, deception, and Lil Miquela. In I. Chiluwa & S. Samoilenko (Eds.), *Handbook of research on deception, fake news, and misinformation online* (pp. 87–103). IGI Global.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., & 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*. 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* (pp. 59–66).

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

Bryson, J. (2010a). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and eesign Issues* (pp. 63–74). John Benjamins Publishing.

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. (2015, August 4). Clones should NOT be slaves. *Adventures in NI*. https://joanna-bryson.blogspot.com/2015/10/clones-should-not-be-slaves.html

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

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

Bryson, J. (2019a, January 4). Bullying and shunning are problems, not solutions. *Adventures in NI*. https://joanna-bryson.blogspot.com/2019/04/bullying-and-shunning-are-problems-not.html

Bryson, J. (2019b, April 7). What we lost when we lost Google ATEAC. *Adventures in NI*. 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. (2011). Just an artifact: Why machines are perceived as moral agents. *Twenty-second international joint conference on artificial intelligence*.

Bryson, J., & Kime, P. (1998). Just another artifact: Ethics and the empirical experience of AI. *Fifteenth international congress on cybernetics* (pp. 385–392).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency* (pp. 77–91).

Carpenter, J. (2016). *Culture and human-robot interaction in militarized spaces: A war story*. Routledge.

Cave, S. (2020, February). The problem with intelligence: Its value-laden history and the future of AI. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 29-35).

Cheng, L. (2018, November 14). Microsoft introduces guidelines for developing responsible conversational AI. *Official Microsoft Blog*. https://blogs.microsoft.com/blog/2018/11/14/microsoft-introduces-guidelines-for-developing-responsible-conversational-ai/

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.

Compton, K., Pagnutti, J., & Whitehead, J. (2017). A shared language for creative communities of artbots. *Proceedings of the 2017 co-creation workshop*.

Crawford, K., & Joler, V. (2018). Anatomy of an AI system-The Amazon Echo as an anatomical map of human labor, data and planetary resources. *AI Now Institute and Share Lab*, *7*.

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. (2017). 'Who's Johnny?': Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, K. Abney, & Jenkins, Ryan (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 173–190). Oxford University Press. https://www.doi.org/10.1093/oso/9780190652951.003.0012

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). Edward Elgar Publishing. 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.

Electronic Frontier Foundation. (2018, May 21). *EFF letter opposing California bot disclosure bill, SB 1001—First Amendment Concerns*. Electronic Frontier Foundation. 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, June 18). *Sophia and her critics*. Medium. https://medium.com/@eripsa/sophia-and-her-critics-5bd22d859b9c

Estrada, D. (2018b). Value alignment, fair play, and the rights of service robots. *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 102–107).

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

Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 57–63).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689–707. 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*. 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. (2004). Oppression. In V. Taylor, N. Whittier, and L. Rupp (Eds.), *Feminist frontiers* (6th ed.). McGraw-Hill Education. (Original work published 1983 in *The politics of reality: Essays in feminist theory*. Crossing Press.)

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

Geraci, R. M. (2006). Spiritual robots: Religion and our scientific view of the natural world. *Theology and Science*, *4*(3), 229-246.

Gibney, E. (2017). How rival bots battled their way to poker supremacy. *Nature News*, *543*(7644), 160. https://doi.org/10.1038/nature.2017.21580

Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., & Crowcroft, J. (2017). Of bots and humans (on twitter). *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 349–354).

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 S. P. van Rysewyk & M. Pontier (Eds.), *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*. 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*. https://arxiv.org/abs/1903.05696

Horsthemke, K. (2017). Biocentrism, ecocentrism, and African modal relationalism: Etieyibo, Metz, and Galgut on animals and African ethics. *Journal of Animal Ethics*, *7*(2), 183-189.

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* (pp. 1311–1320).

Jackson, Z. I. (2020). *Becoming human: Matter and meaning in an antiblack world*. NYU Press.

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

Johnson, B., & Lichfield, G. (2019, April 6). Hey Google, sorry you lost your ethics council, so we made one for you. *MIT Technology Review*. 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. In E. Katz, A. Light, D. & Rothenberg (Eds.), *Beneath the surface: Critical essays in the philosophy of deep ecology* (pp. 17–42). MIT Press.

Kazemi, D. (2013 March 16) Basic Twitter bot etiquette. *Tiny Subversions. http://tinysubversions.com/2013/03/basic-twitter-bot-etiquette/*

Kazemi, D. (2015 June 23) Transphobic joke detection. *Tiny Subversions*. http://tinysubversions.com/notes/transphobic-joke-detection/

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* (p. 1–13). https://doi.org/10.1145/3290605.3300569

Khader, S. J. (2018). *Decolonizing universalism: A transnational feminist ethic*. Springer.

Knight, W. (2019, April 1). *Google employees are lining up to trash Google's AI ethics council*. MIT Technology Review. https://www.technologyreview.com/2019/04/01/1185/googles-ai-council-faces-blowback-over-a-conservative-member/

Konopka, A. (2013). Public, ecological and normative goods: The case of deepwater horizon. *Ethics, Policy & Environment*, *16*(2), 188–207.

Koops, E. (2006). Should ICT regulation be technology-neutral? In B.-J. Koops, C. Prins, M. Schellekens, M. Lips (Eds.), *Starting points for ICT regulation* (pp. 77–108). TMC Asser Press.

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

Latonero, M. (2018). *Governing artificial intelligence: upholding human rights and dignity*. Data&Society. 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. In R. Scharff and Val Dusek, *Philosophy of technology: The technological condition* (pp. 126–137). Blackwell Publishing.

Leopold, A. (1949). *A Sand County almanac, and sketches here and there*. Oxford University Press.

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: Interrupt, inspire, and expose. In A. Nocella II, K. Socha, R. White, & E. Cudworth (Eds.), *Anarchism and animal liberation: Essays on complementary elements of total liberation* (pp. 179–193). McFarland.

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

Metz, T. (2017). Values in China as compared to Africa: Two conceptions of harmony. *Philosophy East and West*, *67*(2), 441–465.

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, National Humanities Center*. https://nationalhumanitiescenter.org/on-the-human/2011/04/political-economy-of-personhood/

Müller, K., & Schaeffer, J. (2018). *Man vs. machine: Challenging human supremacy at chess*. SCB Distributors.

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.

Neff, G., & Nagy, P. (2016). Automation, algorithms, and politics| Talking to Bots: Symbiotic agency and the case of Tay. *International Journal of Communication*, *10*, 4915–4931.

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.

O'Leary, M. (2016) Ethical bot-making. *mewo2.com*. http://mewo2.com/notes/bot-ethics/

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)* (pp. 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). 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

Pierotti, R., & Wildcat, D. (2000). Traditional ecological knowledge: the third alternative. *Ecological Applications*, *10*(5), 1333-1340.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

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.

Richardson, L. (2013, November 27) Bots should punch up. *News You Can Bruise.* https://www.crummy.com/2013/11/27/0

Risse, M. (2018, December 1). Human rights and artificial intelligence: The long (worrisome?) view. *Human Rights, Ethics, and Artificial Intelligence: Challenges for the next 70 Years of the Universal Declaration, The Carr Center for Human Rights Policy.* https://youtu.be/YniwuPWhHSo

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

Roth, Y. [@yoyoel] (2018, November 2) *The same way we sometimes see people dismissing facts as "fake news," we also see real people labeling each other* [Tweet]. Twitter. https://twitter.com/yoyoel/status/1058471834947964928

Romero, S. (2018, December 31). Wielding rocks and knives, Arizonans attack self-driving cars. *New York Times*. 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 Francisco Chronicle*. 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., Laschi, C., Oh, S. R., & Dario, P. (2010). How safe are service robots in urban environments? Bullying a robot. *Proceedings of the 19th international symposium in robot and human interactive communication* (pp. 1–7). https://doi.org/10.1109/ROMAN.2010.5654677

Savage, S., Monroy-Hernandez, A., & Höllerer, T. (2016). Botivist: Calling volunteers to action using online bots. *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 813–822).

Schlesinger, A., O'Hara, K. P., & Taylor, A. S. (2018, April). Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-14).

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

smith, s. e. (2017, October 30). *The ethics of internet bots*. Bitch Media. https://www.bitchmedia.org/article/the-ethics-of-internet-bots

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*, (pp. 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 Press.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A.-K. (2017). Do social bots dream of electric sheep? A categorisation of social media bot accounts. *Proceedings of the 28th Australasian conference on information systems (ACIS)*

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

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* (pp. 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*. MIT Press.

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*. The Verge. 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. In R. Latham (Ed.), *Science fiction Criticism: An Anthology of Essential Writings* (pp. 352–363). Bloomsbury Academic.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & 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. (2018, May 7). What it's like to be a bot. Real Life Magazine. https://reallifemag.com/what-its-like-to-be-a-bot/

Williams, D. (2019a). Consciousness and conscious machines: What's at stake? *AAAI Spring Symposium: Towards Conscious AI Systems*. http://ceur-ws.org/Vol-2287/paper5.pdf

Williams, D. (2019b, June 8). Heavenly bodies: Why it matters that cyborgs have always been about disability, mental health, and marginalization. https://dx.doi.org/10.2139/ssrn.3401342

Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.

Wise, T. (2005, August 13). Animal whites: PETA and the politics of putting things in perspective. *Tim Wise*. http://www.timwise.org/2005/08/animal-whites-peta-and-the-politics-of-putting-things-in-perspective/

Wittkower, D. E. (2020) Privacy as care in the internet of things. In Wiltse, H. (Ed.). *Relating to things: Design, technology and the artificial.* Bloomsbury Publishing.

Wittkower, D. E. (in press). What is it like to be a bot? In S. Vallor (Ed.) *Oxford Handbook of Philosophy of Technology.* Oxford University Press.

Young, I. M. (1988). Five faces of oppression. *Philosophical Forum*, *19*, 270–290.