Race in the Eye of the Robot Beholder: Against Racial Representation, Recognition, and Reasoning in Robotics Research

Tom Williams*
MIRRORLab
Colorado School of Mines
Golden, CO, USA
twilliams@mines.edu

***Abstract*—Significant segments of the HRI literature rely on or promote the ability to reason about human traits like age, gender, and cultural background. In parallel, a significant number of researchers in the Computer Vision community are promoting new methods for (ostensibly) automatically recognizing these types of traits, including race. In this paper, I will argue against the representation, recognition, or reasoning over race by interactive robots, on the basis of ontological, perceptual, and deployment-oriented concerns, and explain why these concerns outweigh other reasonable concerns regarding the risks of colorblindness. Finally, I conclude with a discussion of what this means for the HRI community, and speculate as to possible paths forward.**

I. Introduction

Many social robotics applications depend on robots detecting and recognizing human interactants; and many language-capable robotic applications depend on or assume the ability to refer to and describe humans. Moreover, there has been substantial research interest in the idea of *personalized* robots, that use their model of an interactant to provide an ideal interactive experience for that interactant. The U.S. National Science Foundation's *National Robotics Inititive 3.0*, for example, explicitly calls for work on long-term care robots personalized to individuals. To enable this personalization, researchers have argued that robots need to be able to recognize or otherwise take into account a wide variety of human traits, including norms and dispositions that vary between geographically distinct cultures [1]. One interactant trait that robots *could* try to represent, recognize, or reason over, is interactant race. Indeed, there have been many attempts within the Computer Vision community to try to automatically recognize the race of individuals from camera data[1]. One can naturally anticipate a collision course between these research vectors whereby HRI researchers attempt similar projects in the name of robot personalization or effective language generation.

In this paper, I attempt to warn off researchers from this direction, arguing against the representation, recognition, or reasoning over race by interactive robots, on the basis of ontological, perceptual, and deployment-oriented concerns, and explain why these concerns outweigh other reasonable concerns regarding the risks of colorblindness. Finally, I conclude with a discussion of what this means for the HRI community, and speculate as to possible paths forward.

II. Race: Definition, Origin, and Use

Before discussing the concerns that surround proposals to have robots and other computing technologies recognize, represent, or reason about racial identity, let us first be clear about the core concepts under discussion, which are not typically considered with a critical and precise eye in the HRI literature. Race is a socially constructed [2] structure [3], knowledge system [4], tool [5], or technology [6] for separating a population into hierarchically organized categories or castes (predominantly but not exclusively by those sorted into the dominant category or caste) so that value (and thus power) can be differentially ascribed to these categories or castes. While the origins of race were informed by a variety of religious, social, and political factors (especially the evolution of the English conquest of Ireland over the twelfth through seventeenth centuries [5]), race can be thought of as a fundamentally American invention, with the first modern racial paradigm (with its constituent categories, hierarchy, and politiculture [5]) emerging in the U.S. in the 17th century [7] in order to satisfy the labor requirements of the English capitalist / colonialist projects in the so-called "New World" [8]. The third iteration of this paradigm [5] saw the racialization of those enslaved in these colonialist projects [9] on the basis of primarily cultural and religious (rather than phenotypic) differences [10], for ostensibly religious – but more likely capitalist – reasons [4].

Understanding the origins of the first American racial paradigm is critical not only because they help us to understand what race is and how and why it was created, but also because the *spatially differential transitions* from that seventeenth century paradigm to those used to structure today's

*The author is a White-identifying male American Computer and Cognitive Scientist at a predominantly White and Male American Engineering Institution. This is a relevant caveat that readers should keep in mind while considering his argument.

[1] To avoid further elevating these works I will not cite them herein, but hundreds of such papers can be found by searching for terms like ("computer vision" AND ("race recognition" OR "race classification")) in scholarly search engines.

societies further help us to articulate three key dimensions of race that will be critical in our discussions below.

First, we can consider how the set of racial categories have evolved in the United States. While the official position of the U.S. government as encoded in its census structure is that there are currently five racial categories in the U.S., many sociologists argue that the current U.S. racial paradigm is still essentially binary [11], although some have argued that it is transitioning into a ternary system (whites, honorary whites, and the collective black) [12]. To reiterate, because it is critical for our discussion below, the number of racial categories within a paradigm evolve over time.

Second, we can consider how the makeup of these racial categories have evolved in the United States. Expanding on foundational accounts [13] to consider a wider range of ethnic groups, Treitler carefully describes the successful *ethnic projects* undertaken by many ethnic groups (such as the Irish, Chinese, Italians, and Jews) who, after being initially racialized as (or functionally equivalent to) Black (or, in Bonilla-Silvan terminology, as members of the "Collective Black"), underwent a process of "whitening" in which they were able through concerted effort (typically at the expense of other ethnic groups, especially African Americans), to increase their racial status and change the collective "common sense" as to which racial group they belonged [5]. To reiterate, because it is critical for our discussion below, the mapping from ethnic groups to racial categories evolves over time within a racialized society.

Finally, we can consider the spatial differences in the racial categorization schemes that help constitute different racial paradigms. It can be argued that much of the world (especially the parts of the world acutely influenced by U.S. and British imperialism) shares a *similar* racial paradigm to the U.S., especially its particular White-Black axis, in part due to the wide-ranging influence of American and British imperialism, and more generally due to the origins of all racial categorization schemes in those developed in early America. However, significant differences exist even within heavily anglo-centric cultures, such as the U.S., Australia, and South Africa [14], in part due to the unique ethnic compositions and ethnic projects conducted within these societies. Moreover, starker differences exist in countries with less anglo-centric cultures. Brazil, for example, has been argued to have dozens to hundreds of distinct racial groups [5][2]. And for many African immigrants, there is a common perception of *becoming Black* upon immigrating to America [15]. To reiterate, because it is critical for our discussion below, different countries are structured according to different racial paradigms comprised of different racial categories, and how one racially identifies and/or how one is raced may change as one travels between countries.

With an understanding of the basics of race and racial paradigms, we can begin to discuss concerns that may arise surrounding the development of robots that recognize, represent, or reason over race.

III. Concerns for Roboticists

A. Ontological Concerns

The first set of concerns for roboticists are ontological in nature. Again, the concerns I raise in this paper arise due to trends in the Computer Vision community to attempt to recognize race, and attempts in the HRI community to personalize interactive robot behaviors to dimensions such as culture and gender, leading to an overarching worry that the intersection of these research vectors might lead to roboticists attempting to personalize interactions to interactant "race".

Attempts to do so autonomously, explicitly, and on-the-fly would seem to require robots to explicitly or implicitly assign some racial label to interactants, in order to achieve this tailored personalization. This would in turn require robots to use some representation of a set of racial categories, likely provided by the robot's designer. But committing to any particular system of racial categorization inherently plays into racist logics and would turn a robot into a vehicle through which roboticists would wield race-as-technology or race-as-tool. This can be seen in several ways.

First, if roboticists select and encode a particular set of racial categories into their technologies, they reify and reinforce that categorization scheme as legitimate. Moreover, selecting and encoding any set of racial categories would seem to presuppose that individuals innately have a particular race that can be definitively coded. Similarly, associating a specific racial label with a particular user makes a claim that that specific person objectively falls into that particular racial category.

Second, even if racialization-by-designer-through-robot were deemed beneficial in order to, e.g., interact with users in a way that reaffirms their own likely racial identities, doing so would require constant revision of robots' methods for classification. Unless developers were prepared to periodically reassess the racial categories selected, and unless a robot were prepared to periodically reassess how users were racially categorized, the use of a particular set of racial categories would seem to ignore the dynamic nature of those categories, and the way that racial categories (used by humans living in racialized societies to racialize themselves and others) change over time, both in terms of the set of categories that are used within a racial paradigm, how those categories are hierarchically arranged, and how ethnicities are mapped to or associated with those categories. And even if roboticists were prepared to enact these continuous changes, this would be an explicitly racist act conducted under racist logics, and would be especially troubling given the current status quo in racialized societies like the United States, in which those empowered to make these decisions are predominantly racialized as White.

Finally, if a robot is designed to use a particular set of racial categories, (temporarily ignoring the arguments above as to why this would be a bad idea), if that robot or its software architecture is made available for use in other countries, this would seem to ignore the fact that different countries

[2] See Ch.2 Fn.27 for a discussions of the complexities of racial category enumeration in Brazil.

use different racial categorization schemes. If a robot were designed to categorize users according to the set of racial categories used in the U.S., for example, and that robot were then deployed in another country, this would serve as a vehicle for propagating and globalizing the U.S.'s system of racial categories, and could be seen as part of the larger colonialist and white supremacist projects enacted by the U.S., or more generally as part of the "transnational assemblage" of racist logics [16]. Moreover, when a robot classifies an individual user according to a particular scheme, unless the robot is prepared to reassess how that user is classified when either the robot or the user are re-located to a different country, this would seem to ignore the spatial dynamics of racial categorization. More generally, efforts to classify individuals according to race and then store that information as a static user trait ignore the nature of racial categorization as a dynamic phenomena that is always performed from a particular spatiotemporosocial perspective. I highlight these issues to note the fundamental infeasibility and impracticality of robots storing and using labels of interactant race; but I would once again stress that even if robotic assignment of racial categories to users were not infeasible or impractical on these grounds, attempting to do so would reflect a design perspective grounded in racist logics in which robots would serve as a means for designers to wield race-as-technology or race-as-tool.

To provide a demonstrative example of these problems, we can consider Microsoft's MS-CELEB-1M dataset. As detailed by Scheurman et al. [17] in their examination of racial categorization in computer vision databases, Microsoft chose to label the faces in that dataset with the categories "Caucasian," "Mongoloid," and "Negroid," on the basis that these categories encompassed "all the major races in the world". By doing so, Microsoft reinforced several intersecting notions: (1) that everyone in the world can be assigned to a consistent set of racial categories; (2) that a single set of racial categories are universally applicable; (3) that those three categories are the categories that are universally used to categorize people; and (4) that race is a scientific or biological rather than social concept (a presupposition tied to those particular terms). Moreover, by labeling individual faces according to this categorization scheme, Microsoft implicitly claimed that the individuals in their datasets should be and are socially classified according to that scheme. And finally, by using these labels in their dataset, Microsoft researchers implicitly provided others with a computationally-augmented opportunity to use race-as-technology, in a way that would propogate their own particular racist logics and worldview. Roboticists classifying interactant race using models trained on this dataset would implicitly buy into these claims and wield race-as-technology on behalf of those who constructed the dataset. But moreover, due to the unique persuasive power that robots wield, if this classification were communicated by robots in any way, then roboticists would risk actively reinforcing these notions in the minds of interactants.

B. Perceptual Concerns

In the previous section, I argued that having robots use a particular racial categorization scheme, and assigning racial categories to individuals within a robot's memory, is a design perspective that makes, reifies, and reinforces problematic and fallacious claims and notions, and which ignores the spatiotemporosocial dynamics and the very nature and history of race. In this section, I will consider how racial categories would be associated with interactants in a robot's memory in the first place.

First, we can consider the difference between racialization and racial identity. An individual's racial identity and how they are racialized within a given social system may necessarily differ due to the spatial and social dynamics of race described above. This means that whether robot-internalized racial categorizations originate from self-reports from interactants versus automatic categorization, for example, via machine learning classifier, necessarily asserts the primacy of one source or the other. Collecting racial identity or performing perceptual racialization would both be concerning, however, as both could be viewed as a form of biometric surveillance. In her book *Dark Matters* [18], for example, Browne describes the history of racialized surveillance technologies, drawing connections from the 16th century Book of Negroes through to more modern digital surveillance technologies such as Databases. Mobile, perceptual, agentic technologies like robots can be seen as a further extension of this trend, regardless of the source of the labels encoded in a robot's memories.

Moreover, racial categorization of interactants on the basis of perceptual data again requires adherence to fallacious and racist logics. Automated racial categorization on the basis of camera data can be seen as a form of *digital epidermalization* [19], whereby the categorizing technology races and racializes, imposing race onto the body observed. This is problematic in and of itself for multiple reasons. First, digital epidermalization reifies and reinforces in-built, perspectiveless, and spatiotemporally static notions of race, as described above. Second and relatedly, digital epidermalization reinforces fallacious notion of essential differences; a process Scheuerman et al. refer to as auto-essentialization [20]. Third, digital epidermalization races and racializes without the consent of the observed, and (assuming that visual classification is used in any meaningful way) forces that racialization to be acknowledged and accepted by others. Fourth, because face detection, recognition, and classification technologies tend to work poorly for people of color (especially women of color), digital epidermalization privileges whiteness (and white maleness) [21]. And finally, digital epidermalization fundamentally (and falsely) asserts that race is something that can be objectively perceived through visual stimuli, reasserting problematic and fallacious equivalences between race and visually discernible, phenotypic markers such as skin color.

Finally, it is worth noting that these concerns regarding digital epidermalization and the perception of race arise regardless of the provenance and annotation of the data used to effect automated race classification. To re-consider the example used in the previous section, even had Microsoft contacted those whose images were stored in their dataset and solicited how those people racially identified (rather than, as one might assume they did, asking crowdworkers to provide these labels), as soon as those self-reported racial identities were used to train a predictive model (e.g. for deployment on robotic platforms), the resulting model would nonetheless be subject to these concerns.

C. Deployment Concerns

Finally, regardless of what racial ontology a robot might use, and how a robot might categorize people according to that ontology, and regardless of whether or not such efforts would inherently operate according to and reinforce racist logics, there are fundamental concerns about racialized perception and data technologies that transcend model parameterizations or data bias concerns. That is, one might fundamentally ask why one is trying to perceive and store race in the first place, who has access to this data, what they might use it for, and how this might shift power in inequitable ways: questions raised by justice-oriented "third-wave" AI Ethics frameworks [22]. Like all racialized surveillance technologies, race-classifying robots would present opportunities for the persecution and oppression of minoritized racialized groups [23], [24], especially by institutions that were created for this purpose or which have historically sought such goals, such as law enforcement agencies [25]. Indeed, researchers have raised concerns about the use of many of the technologies used in the HRI community that rely on face detection [23] due to the potential for law enforcement to use these technologies to systematically oppress people of color (see also [26]), extending existing centuries-long trends of using surveillance technologies as a tool of racial oppression [18], [27], [28]) These concerns are especially relevant and troubling given existing trends in the robotics community regarding the development of robots for the police[3].

IV. Counterarguments and Paths Forward

Thus far I have presented a variety of arguments against the development of robots that purport to represent, recognize, or reason over race. One reasonable counterargument would be to raise concerns regarding colorblindness. As numerous scholars have pointed out, adherence to a colorblind ideology in which one refuses to recognize or consider the *social reality* of race is itself a form of racism, which perpetuates the racial status quo through studied ignorance [29]. For language-capable robots in particular, one could argue that failure to recognize how someone is likely to be racialized could mean an inability to reaffirm users' racial identities, and an inability to recognize and respond to racialized microaggressions.

Recent work in the HRI literature has suggested that robots have the potential to exert moral influence on human interactants, and that a failure to recognize and respond to norm violating requests could be viewed as tacit acceptance [30]. Similar research has argued that the typical design objectives of the HRI community could lead to designing robots that inadvertently lean into overt and benevolent sexism [31]. Other recent work has thus argued for the creation of robots through an explicitly Feminist design stance, in which intentionally female-presenting robots push back on overt or benign sexism in ways that subvert (socially harmful) gender norms and expectations [32], [33]. One can imagine a similar argument being made for explicitly anti-racist robots, that refuse to accept commands to perform overt or implicitly racist acts, that recognize and call out racist microaggressions, and that can actively uplift and celebrate historically oppressed racialized groups [34]. Enabling robots to ascertain the likely racial identity of interactants could facilitate such approaches.

To reason through the merits of this counterargument, we can consider insights from closely related fields. First, some researchers in the Computer Vision and HCI communities have argued that representing racial identity data can be appropriate *if* users willingly share this information [35]. This could suggest that if users are providing this data rather than attempting to classify it from perceptual data, this could be appropriate. But this may only be justifiable if this data is being used in a way similar to how it is used in that photo captioning work, e.g., to describe one's racial identity to others. It is not clear whether this would also be acceptable for the purposes of robot behavior personalization and automated adaptation. Moreover, in recent work, Bennett et al. [35] present perspectives from populations under-represented in HCI, who strongly argued for image captioning systems to err on the side of less politicized descriptions of appearance, especially when self-identifying information could not be maintained. This further underscores the need for robots to avoid attempting to recognize, represent, and reason over race in most cases. In fact, this could provide an opportunity for robots to positively wield their persuasive power, by setting an example and *not* relying on this type of information in cases where it has not been provided and encouraged by a human referent. Even in those cases, however, we would urge caution because (as described above) this would nevertheless further turn robots into racialized surveillance technologies.

A promising conceptual path forward might be found by considering other areas of the ethical robot design literature. Moor [36], for example, recommends distinguishing between explicit ethical agents (robots designed to explicitly reason over ethical principles) vs implicit ethical agents (robots whose actions are constrained in ways that help prevent unethical actions from being taken, without providing explicit reasoning capabilities). In the robot ethics literature, researchers have used this framework to argue for the use of explicit ethical agents in various contexts [37]. However, this same framework could similarly be used to argue for explicitly designing for *implicit ethical agency* in use contexts where concerns

[3] As above, I am choosing not to cite these works explicitly to avoid further elevating these works, but many papers on this topic can be found by searching terms like (("law enforcement" OR "police") AND "robot"). The author would also acknowledge here that he is one of the lead petitioners in the No Justice No Robots campaign and thus has publicly committed to advocacy against such robotics projects.

surrounding autonomous moral agents arise [38]. Just as the domains in which explicit ethical agents can be deployed may be limited, so too may the domains in which explicitly race-aware agents can be deployed may be narrow (e.g., to domains where interactants can provide their racial identities, where this information does not need to be stored in association with other personally identifiying data, and where these identities are used in the context of conversations where this is deemed by those interactants to be important and acceptable). In other contexts, it may be better to only design *implicit* race-aware agents, wherein race is considered as a design factor (especially in contexts where robots are being designed by and for historically excluded populations [34] (see also [39]), but is not explicitly recognized, represented, or reasoned over by robots – or rather, by roboticists, through their robots.

V. Conclusion

In this paper, I briefly considered the myriad risks of attempting to recognize, represent, and/or reason about race in interactive robotic systems. While roboticists should be cognizant of the dangers of colorblindness, and of the utility of *implicitly* racializing robots, I have argued that in most cases roboticists should refrain from explicitly computationally recognizing, representing, or reasoning over race in their work. I hope that this paper brings awareness to these risks, and helps roboticists to steer away from those racialized robotics technologies I have argued that our field seems to be approaching. Finally, I hope that this paper will encourage roboticists to think more critically about the overarching implications of what they choose to recognize, represent, and reason over in their efforts to achieve robotics design goals such as personalization.

References

[1] N. Gasteiger, M. Hellou, and H. S. Ahn, “Factors for personalization and localization to optimize human–robot interaction: A literature review,” *International Journal of Social Robotics*, pp. 1–13, 2021.

[2] A. Smedley and B. D. Smedley, “Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race.” *American psychologist*, vol. 60, no. 1, p. 16, 2005.

[3] E. Bonilla-Silva, “Rethinking racism: Toward a structural interpretation,” *American sociological review*, pp. 465–480, 1997.

[4] A. Smedley and B. D. Smedley, *Race in North America: Origin and evolution of a worldview*. Westview Press, 2012.

[5] V. B. Treitler, *The Ethnic Project*. Stanford University Press, 2020.

[6] B. Coleman, “Race as technology,” *Camera obscura: feminism, culture, and media studies*, vol. 24, no. 1, pp. 177–207, 2009.

[7] A. Smedley, “” race” and the construction of human identity,” *American Anthropologist*, vol. 100, no. 3, pp. 690–702, 1998.

[8] ——, “Antecedents of the racial worldview,” *Race and racilization: Essential readings*, pp. 31–44, 2007.

[9] M. Omi and H. Winant, “On the theoretical status of the concept of race,” *Race, identity and representation in education*, pp. 3–10, 1993.

[10] A. S. Parent, *Foul Means: The Formation of a Slave Society in Virginia, 1660-1740*. UNC Press Books, 2003.

[11] J. R. Feagin, *Racist America: Roots, current realities, and future reparations*. Routledge, 2000.

[12] E. Bonilla-Silva, “From bi-racial to tri-racial: Towards a new system of racial stratification in the usa,” *Ethnic and racial studies*, vol. 27, no. 6, pp. 931–950, 2004.

[13] N. Ignatiev, *How the Irish became white*. Routledge, 2012.

[14] K. Farquharson, “Racial categories in three nations: Australia, south africa and the united states,” in *Proceedings of ‘Public sociologies: lessons and trans-Tasman Comparisons’, the Annual Conference of The Australian Sociological Association (TASA)*, 2007.

[15] I. Wilkerson, *Caste: The origins of our discontents*. Random House, 2020.

[16] M. Christian, “A global critical race and racism framework: Racial entanglements and deep and malleable whiteness,” *Sociology of Race and Ethnicity*, vol. 5, no. 2, pp. 169–185, 2019.

[17] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker, “How we’ve taught algorithms to see identity: constructing race and gender in image databases for facial analysis,” *Proceedings of the ACM on Human-computer Interaction*, vol. 4, no. CSCW1, pp. 1–35, 2020.

[18] S. Browne, *Dark matters*. Duke University Press, 2015.

[19] ——, “Digital epidermalization: Race, identity and biometrics,” *Critical Sociology*, vol. 36, no. 1, pp. 131–150, 2010.

[20] M. K. Scheuerman, M. Pape, and A. Hanna, “Auto-essentialization: Gender in automated facial analysis as extended colonial project,” *Big Data & Society*, vol. 8, no. 2, p. 20539517211053712, 2021.

[21] J. A. Buolamwini, “Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers,” Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[22] M. Le Bui and S. U. Noble, “We’re missing a moral framework of justice in artificial intelligence,” in *The Oxford Handbook of Ethics of AI*. Oxford University Press, 2020, p. 163.

[23] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. Sánchez *et al.*, “Ai now 2019 report. new york: Ai now institute,” 2019.

[24] M. L. Littman, I. Ajunwa, G. Berger, C. Boutilier, M. Currie, F. Doshi-Velez, G. Hadfield, M. C. Horowitz, C. Isbell, H. Kitano, K. Levy, T. Lyons, M. Mitchell, J. Shah, S. Sloman, S. Vallor, , and T. Walsh., “Gathering strength, gathering storms: The one hundred year study on artificial intelligence (ai100) 2021 study panel report,” http://ai100.stanford.edu/2021-report, September 2021.

[25] A. S. Vitale, *The end of policing*. Verso Books, 2017.

[26] C. Garvie, *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

[27] R. Benjamin, “Race after technology: Abolitionist tools for the new jim code,” *Social Forces*, 2019.

[28] M. Alexander, *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2012.

[29] E. Bonilla-Silva, *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers, 2006.

[30] R. B. Jackson and T. Williams, “Language-capable robots may inadvertently weaken human moral norms,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 401–410.

[31] R. B. Jackson, T. Williams, and N. Smith, “Exploring the role of gender in perceptions of robotic noncompliance,” in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 559–567.

[32] K. Winkle, G. I. Melsión, D. McMillan, and I. Leite, “Boosting robot credibility and challenging gender norms in responding to abusive behaviour: a case for feminist robots,” in *Companion of the 2021 ACM/IEEE international conference on human-robot interaction*, 2021, pp. 29–37.

[33] K. Winkle, R. B. Jackson, G. I. Melsión, D. Bršcic, I. Leite, and T. Williams, “Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study,” in *Proceedings of the 2022 ACM/IEEE international conference on human-robot interaction*, 2022.

[34] J. Nias, L. Hampton, P. Sampson, and M. Ruffin, “Decolonizing technologies for preserving cultural and societal diversity,” in *CHI 2020 Workshop on Engaging in Race in HCI*, 2020.

[35] C. L. Bennett, C. Gleason, M. K. Scheuerman, J. P. Bigham, A. Guo, and A. To, ““it’s complicated”: Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.

[36] J. H. Moor, “The nature, importance, and difficulty of machine ethics,” *IEEE intelligent systems*, vol. 21, no. 4, pp. 18–21, 2006.

[37] M. Scheutz, “The case for explicit ethical agents,” *Ai Magazine*, vol. 38, no. 4, pp. 57–64, 2017.

[38] A. Van Wynsberghe and S. Robbins, “Critiquing the reasons for making artificial moral agents,” *Science and engineering ethics*, vol. 25, no. 3, pp. 719–735, 2019.

[39] Y. A. Rankin and I. Irish, “A seat at the table: Black feminist thought as a critical framework for inclusive game design,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–26, 2020.