The Eye of the Robot Beholder: Ethical Risks of Representation, Recognition, and Reasoning over Identity Characteristics in Human-Robot Interaction

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 identity characteristics, including age, gender, and cultural background. However, attempting to handle identity characteristics raises a number of critical ethical concerns, especially given the spatiotemporal dynamics of these characteristics. In this paper I question whether human identity characteristics can and should be represented, recognized, or reasoned about by robots, with special attention paid to the construct of race, due to its relative lack of consideration within the HRI community. As I will argue, while there are a number of well-warranted reasons why HRI researchers might want to enable robotic consideration of identity characteristics, these reasons are outweighed by a number of key ontological, perceptual, and deployment-oriented concerns. This argument raises troubling questions as to whether robots should even be able to understand or generate descriptions of people, and how they would do so while avoiding these ethical concerns. Finally, I conclude with a discussion of what this means for the HRI community, in terms of both algorithm and robot design, and speculate as to possible paths forward.

CCS CONCEPTS

• **Social and professional topics → Race and ethnicity**; *Gender*; *Cultural characteristics*; Surveillance; • **Computer systems organization → Robotics**.

KEYWORDS

Race, Gender, Ethics, Identity, Personalization

ACM Reference Format:
Tom Williams. 2023. The Eye of the Robot Beholder: Ethical Risks of Representation, Recognition, and Reasoning over Identity Characteristics in Human-Robot Interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3568294.3580031

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580031

1 INTRODUCTION

The theme of this year's Human-Robot Interaction conference is "HRI for All". What does this mean? It could mean making sure that our robot designs are not implicitly (or explicitly) centering the needs, values, perspectives, and desires of wealthy, white, non-disabled heterosexual men. Alternatively, taking a subtly different turn, it could mean making sure that our robot designs are explicitly designed with, for, and to meet the needs, values, perspectives, and desires of, non-wealthy, non-white, disabled, non-heterosexual, and/or non-masculine or non-male interactants and users. These are worthwhile goals that our community should absolutely pursue; yet the approach towards these goals can take a variety of shapes.

Even taking an explicitly community-oriented participatory design perspective (e.g., Design Justice [24] or Engineering Justice [53]), the ways that communities' needs, values, perspectives and desires are translated and manifested into robot designs can vary drastically, with enormous ethical implications. This approach to equitable robot design might manifest in at least three ways: (1) as an entirely new robot product designed by, with, and for a particular community; (2) as a robot whose appearance and behavior can be changed by members of such a community; (3) or, and as discussed in this paper, as a robot whose appearance or behavior change based on the community to which an interactant appears to belong.

This last approach, which one might call *identity-based personalization*, has been popular in communities like HRI, perhaps because it presents distinct computational challenges that computer scientists feel well suited to work on, and are incentivized to work on by academic and governmental systems and structures. Within this area, researchers have suggested personalizing robots' behaviors to a wide variety of factors, including gender, culture, ethnicity, and race [40]. Researchers have argued that robots should be able to adapt to the gender of their interactants in order to demonstrate a more nuanced understanding of their environment [76], be more user friendly [57, 71], better target them with advertisements [34], and better emulate gender differences in behaviors like handshaking [67]. Researchers have argued that robots should be able to adapt to the culture of their interactant, in order to account for differences in norms and dispositions between geographically distinct cultures [39]. And researchers have argued that robots should be able to adapt to the race or ethnicity of their interactant, in order to provide better medical outcomes [96]. While work on identity-based personalization in human-robot interaction may be approached with good intentions, these approaches typically require robots to represent, recognize, and/or reason about these identity characteristics. And in fact, these calls for increased personalization are being raised at a time when related areas like the Computer Vision community are seeing dramatic upticks in attempts to automatically recognize the race, ethnicity, or religious group of individuals from camera data [29][1].

Similarly, even outside the domain of identity-based personalization, there has been a wealth of recent research that does not explicitly call for representing, recognizing, or reasoning about these characteristics, but nevertheless calls for the development of behaviors for which those capabilities might be natural antecedents. For example, because of observations that robots' failure to call out or explicitly reject unethical behaviors could lead to accidental condoning of those behaviors and a weakening of moral norms [43], researchers have been considering how robots could most effectively call out and reject both overtly hostile racism or sexism [103, 106], or racist or sexist microaggressions [49].

Work being conducted through these perspectives serves to disrupt the status quo logics of robot design by operating through the lens of pro-social justice design frameworks such as Feminist Design [105]. Yet even this work must be pursued with great care, as certain approaches towards such aims may inherently require detection of racist or sexist language, which may in turn require detecting, representing, or reasoning about the race, sex, or gender of the victim of the offending language.

Finally, it is worth noting that there is a vast literature on robotic understanding and generation of referring expressions. While most of this work is focused on understanding and generating descriptions of objects and locations, description of people is a popular use case in this domain [27, 100–102]. Understanding and generating descriptions of people naturally entails understanding and generating identity-based pronouns, adjectives, and nouns, presumably on the basis of some perceptions, representations, and/or reasoning processes, and indeed, HRI researchers have specifically argued for perceiving gender for these purposes [77].

In short, opportunities and requirements to recognize, represent, and reason over human identity characteristics are commonplace throughout the HRI literature, both within and beyond the scope of identity-based personalization.

In this paper, I will argue that while personalization and design efforts that respect and work to the benefit of minoritized identity groups are noble and worthwhile in purpose, these efforts must be approached in a way that avoids representing, recognizing, or reasoning over these identity characteristics, due to a number of key ontological, perceptual, and deployment-oriented concerns. To make this argument, I will specifically consider the example of race. I will begin by considering sociological theories of race and ethnicity. Next, I will discuss how those sociological theories suggest key ontological, perceptual, and deployment-oriented concerns regarding the computational representation, recognition, and reasoning over race. Then, I will describe how similar concerns similarly arise for other identity characteristics such as gender and culture. Finally, I will identify and respond to possible counter-arguments before concluding with a discussion of what this all means for the HRI community and possible productive paths forward.

2 BACKGROUND

2.1 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 [86] structure [14], knowledge system [87], tool [92], or technology [23] for separating a population into hierarchically organized categories[2] (predominantly but not exclusively by those sorted into the dominant category) so that value (and thus power) can be differentially ascribed to these categories. 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 [92]), race can be thought of as a fundamentally American invention, with the first modern racial paradigm (with its constituent categories, hierarchy, and politiculture [92]) emerging in the U.S. in the 17th century [84] in order to satisfy the labor requirements of the English capitalist / colonialist projects in the so-called "New World" [85]. The third iteration of this paradigm [92] saw the racialization of those enslaved in these colonialist projects [65] on the basis of primarily cultural and religious (rather than phenotypic) differences [69], for ostensibly religious – but more likely capitalist – reasons [87].

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 understanding the *spatially differential transitions* from that seventeenth century paradigm to those used to structure today's societies.

2.2 The Temporal Dynamics of Race

The delineation and composition of racial categories used in the United States has evolved (and continued to evolve) over the past four centuries. Legally, categorization schemes in the United States transitioning from *{white, Negro}* to *{white, Black, Asian, Hispanic, "other"}* [92]. Similarly, from a different level of analysis, many sociologists argue that the current U.S. racial paradigm is still in the process of transitioning from an essentially binary system [32] into a ternary system (whites, "honorary whites", and the "collective black") [15].

In parallel, the ethnic composition of these categories has similarly evolved. Expanding on foundational accounts [42] 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 [92].

[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.

[2] Some recent popular accounts have analyzed these categories through the lens of caste [99]. While this is a revealing mode of analysis it is subject to longstanding critiques that go beyond the scope of this paper [25].

To summarize, the set of racial categories and the mapping from ethnic groups to racial categories evolves over time within a racialized society.

2.3 The Spatial Dynamics of Race

To understand the challenges that race poses for the global human-robot interaction community, we must also understand the *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 [31], in part due to the unique ethnic compositions and ethnic projects conducted within these societies. For example, South Africa's racial paradigm has been categorized as *{black, white, colored, Asian}*, and Australia's has been categorized as *{white, black, Asian, Indian, "ethnic looking"}* [31, 36, as cited in [92]].

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 [92][3]. And for many African immigrants, there is a common perception of *becoming Black* upon immigrating to America [99].

To summarize, 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.

2.4 Race in Robotics

Compared to topics like Gender, Race has received relatively little attention in the human-robot interaction community. There have been a small number of articles attributing race to robots or examining how racial prejudices carry over into robotics [10, 20, 30, 46, 88–91], many of which have highlighted the problems with robots being predominantly designed by, for, and in the image of white men. In response to these concerns, scholars like Riek and Howard [73] have called for more diversity in robot morphology and behavior. Similarly, researchers such as Ostrowski et al. [68] have called for more attention to race when considering who gets to design robots, and there have been a few efforts to explicitly design robots with members of minoritized racial [64] and ethnic [59] groups. Similarly, there have been field-wide calls to action against racism [95] and against racist uses of robots [1].

While this small body of work is a welcome start within our community, it does not come close in scope to topics like gender. Even within just the past year, there has been an astounding amount of new research on robots and gender. A recent review by Perugia and Lisy [70] found hundreds of papers on this topic, and a google scholar search for "robot gender" yields over 150 papers on this topic from 2022 alone.

Given the centrality of race within human society, given the increased interest in identity-based personalization within robotics, and given the relative lack of research on race in human-robot interaction, I argue that it is critical to more carefully consider the concerns that race presents for roboticists, especially the concerns that may arise surrounding the development of robots that recognize, represent, or reason over race. As such, in this work I will explicitly consider race as an archtypical-yet-understudied point of analysis within the space of identity-based robot personalization. In the next section I will thus explore three categories of concerns surrounding race in robotics from this perspective.

3 CONCERNS FOR ROBOTICISTS

3.1 Ontological Concerns

The first set of concerns for roboticists are ontological in nature. Attempts to computationally leverage notions of race (especially in a manner that is autonomous, explicit, 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

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

ignore the fact that different countries 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 [21]. 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, let us consider Microsoft's MS-CELEB-1M dataset. As detailed by Scheuerman et al. [82] 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 propagate 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.

3.2 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.

To begin, it is important to distinguish 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* [17], 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 (just as automatic gender recognition systems erroneously treat gender as binary, immutable, and physiological [47]). Second and relatedly, digital epidermalization reinforces fallacious notion of essential differences; a process Scheuerman et al. refer to as auto-essentialization [80]. 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) [19]. Fifth, digital epidermalization's privileging of whiteness may be most pronounced for those most "clearly" deemed white. Just as misgendering serves as a form of structural violence that negatively impacts trans individuals [6, 61] (see also [47]), leading to overwhelming negative perceptions of automatic gender recognition among trans individuals [41], it is reason to expect that automatic race recognition would enact unique structural violence on, and be particularly negatively perceived by, those whose race is not clearly, cleanly, or "accurately" assigned by a robot. And lastly, 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.

3.3 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 [52]. Like all racialized surveillance technologies, race-classifying robots would present opportunities for the persecution and oppression of minoritized racialized groups [26, 58], especially by institutions that were created for this purpose or which have historically sought such goals, such as law enforcement agencies [97]. Indeed, researchers have raised concerns about the use of many of the technologies used in the HRI community that rely on face detection [26] due to the potential for law enforcement to use these technologies to systematically oppress people of color (see also [38]), extending existing centuries-long trends of using surveillance technologies as a tool of racial oppression [3, 11, 18]) These concerns are especially relevant and troubling given existing trends in the robotics community regarding the development of robots for the police[4].

4 GENERALIZATION TO OTHER IDENTITY CONCERNS

While in this paper I have been primarily focusing on race as an organizing example, most of the concerns described thus far also apply to dimensions of identity discussed in the first section, like culture and gender. While similar arguments may be made for human identity characteristics such as class, ability, and sexuality, and the intersections therebetween, I focus on culture and gender in this section for two reasons. First, there has been substantial past work attending to identity-based personalization of robots based on these culture and gender. Second, there have been critically important and timely calls made of late for roboticists to more intentionally integrate cultural and gender-based considerations into their design processes.

4.1 Ontological Concerns

The ontological concerns described above are most straightforwardly re-applicable to gender. While the risks of internally categorizing someone as a man or a woman may not have precisely the same consequences as assigning a racial category, committing to a particular categorical gender categorization system nevertheless presents risks of falsely presupposing a binary or otherwise over-discretized conceptualization of gender. While gender may not have precisely the same spatiotemporal dynamics as race, robotic automatic gender recognition nevertheless run the risk of assuming an immutable. As Keyes [47] points out in their study of the HCI literature, there is a persistent trend in even that body of work to assume that gender is binary and immutable, even in papers that focus on gender; assumptions that deviate from sociological theories of gender [55]. Just as Microsoft's classification of faces as "Caucasian", "Mongoloid", and "Negroid" constituted a commitment to a particular set of political, philosophical, and sociological claims, so too do HRI researchers make similar claims about gender when they commit to a particular gender categorization scheme.

The application of the concerns raised in this paper to culture is less straightforward, in part due to the lack of clarity or agreement on how cultures are proposed to be delineated and discretized in much work on this topic [56, 75], with some roboticists discussing cultures in terms of nationality [9, 66, 93, 98, 103] and/or as "Eastern vs Western" [22] or "Arabic vs Western" [4, 5], and others differentiating between individuals' specific cultural orientations (e.g., individualist vs collectivist [35, 50, 54, 60, 78]). Nevertheless, there is an obvious (if nuanced, confusing, and contentious) relationship between ethnicity and culture, which are (for better or worse) used synonymously at certain levels of cultural, social, and psychological analysis [13] that warrants care, forethought, and precision for those considering the development of culturally adaptive robots. And in general, it seems clear that concerns similar to those discussed above may arise for any of these cultural categorization schemes, if such schemes are used to computationally sort, label, and characterize users.

4.2 Perceptual Concerns

The perceptual concerns described above apply to both gender and culture. First, computationally collecting or perceiving interactant gender or culture could, depending on categorization scheme, be viewed as a surveillance project. Second, computational perception of gender or culture could reinforce erroneously immutable and physiological notions of gender and culture. Third, computational perception of gender and culture similarly risk assignment of categories without consent, and potentially used in ways that require others to acknowledge and accept these non-consenting assignments. Fourth, these approaches are similarly likely to privilege maleness and membership in dominant hegemonic cultural hegemonic groups due to cross-group performance differences. Fifth, as discussed above, these approaches are likely to inflict structural violence towards those on the boundaries of whatever ontological categorization schemes are used for automated perception [24, 47]. And finally, as above, and especially for gender, these approaches would falsely assert that social identity can be confidently and meaningfully perceived through visual stimuli.

[4] 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"). I also acknowledge here that I am one of the lead organizers of the No Justice No Robots campaign and thus have publicly committed to advocacy against such robotics projects.

4.3 Deployment Concerns

Finally, as above, representing, perceiving, and reasoning over gender and culture present ethical concerns when systems move from the lab into the field, where researchers no longer have control over how this data is being used, who as access to the data, and how these data privacy and data use concerns might shift power in inequitable ways. Culture, when viewed from the perspective of a particular location with a particular racialized social system, is wrapped up with race in complex and nuanced ways, regardless of whether culture is viewed from that perspective in terms of nations, broad categories, or personal orientation [7, 33]. Similarly, there have been recent calls across multiple fields for increasing attention to the ways that surveillance is gendered just as it is racialized [2, 8, 28, 51, 62]. Finally, surveillance-wielding organizations like law enforcement groups are known to wield violence in ways that affect people with intersectionally oppressed identities in acute ways [74].

5 COUNTERARGUMENTS

Now that I have discussed the three classes of concerns that recognition, representation, and reasoning over identity characteristics such as race would present for roboticists, I will consider the possible counterarguments that could be made and how those counterarguments may in turn be countered..

First counterargument: Failure to recognize race perpetuates colorblind ideology.

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 [16]. However, we do not expect or desire other pieces of technology (including other language-capable technologies) to recognize, represent, or reason over participant race. And moreover, I argue that colorblindness is primarily a concern for *roboticists* rather than a concern for *robots*. That is, roboticists' recognition of the social reality of race and the ways it needs to be accounted for during the design process does not entail an obligation to computationally model this recognition in robotic technologies.

Second counterargument: Failure to recognize race precludes appropriate recognition and response to racist norm violations.

Concerns over colorblindness motivate a second argument, due to the reasons why colorblind ideology is problematic in human-human interactions. For language-capable robots in particular, one could argue that failure to recognize how someone is likely to be racialized (or gendered) 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 [43]. 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 [45]. 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 [103, 106] (see also [37]). 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 [64]. Enabling robots to ascertain the likely racial identity of interactants could facilitate such approaches.

Considering these potential benefits requires a utilitarian analysis of costs and benefits. It is true that these capabilities would benefit, in some cases, from recognition, representation, and reasoning over a potential victim's identity characteristics. But it is unclear whether the prevalence and severity of this category of micro-aggression outweigh the concerns raised above.

Third counterargument: Representation of race can be grounded in self-identification.

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 [12]. 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. [12] 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, I would urge caution because (as described above) this would nevertheless further turn robots into racialized surveillance technologies. Moreover, because identity characteristics are spatially and temporally dynamic, even relying on self-identification might prove problematic as identification changes over time or with re-location of human or robot.

Finally, while robots in many domains may need to store verbally provided information about other personal preferences or private information, such information cannot be used to justify collection of racial identity labels in the same way. Not only is it unclear for what purpose racial identity labels would be used by robots, but moreover, to use self-provided labels in any meaningful and responsible way would require a nuanced understanding of race and the way that categorization label fits into the broader racialized social system in which the robot is deployed, thus undermining the goal of relyiing only on self-provided labels.

Fourth counterargument: Robotic representations of race need not be temporally static.

In response to my concern about the temporal dynamics of race, one could argue that databases of self-identifications could be periodically revisited, revised, or retracted, in a way similar to computer vision database practices promoted by researchers like Scheuerman et al. [82], through a predesigned data maintenance plan [79]. However, such an approach would seem to require substantial uncompensated and potentially unreasonable effort and labor from those needing to self-report their identities that would be exacerbated by the situated, mobile nature of robots (in contrast to the image captioning contexts considered by Scheuerman et al. [82]). Moreover, given that robotic systems will increasingly involve hosts of spatially distributed yet integrated robots (as opposed to single, monolith online systems), even a system allowing for self-reporting of identity could have worrying surveillance implications. Finally, I would direct the reader to the work of Scheuerman et al. [81] and Hamidi et al. [41] for consideration of other concerns that may arise even in systems allowing for self-reporting of identity.

Fifth counterargument: Users may correct or "opt out" from categorization.
Finally, drawing on Hamidi et al. [41]'s work outside robotics contexts, one could argue that if users are told how they are categorized they could provide corrections to the robot's categorization, or simply opt out at that point. However, it is unclear whether this would really be possible to do effectively in human-robot interactions relying on synchronous verbal communication, especially given privacy concerns that would arise in these cases, and given the substantial potential for psychological and emotional harm from verbal misracializing and misgendering by robots, (especially in multi-user contexts) and the ways that this would push extraneous and potentially painful labor onto human interactants.

6 IMPLICATIONS AND PATHS FORWARD

Given the argument I have laid out in this paper and its counter-counter-arguments, I will now describe some implications for the field of robotics, and possible design paths forward.

First, my overall argument should emphasize to the reader that there are good reasons to attend to identity factors like race in robotics. Yes, robot designers should de-centering whiteness and other overly centered dimensions of identity. Yes, robot designers should avoid colorblindness. Yes, robot designers should enable robotic applications that recognize and celebrate minoritized racial identities. Yes, robot designers should understand the implications of race, gender, and culture in robotics. And eliminating the implicit systems of power such as race, gender, disability and class that underlie formal power structures requires recognizing and interrogating those power structures [48].

However, I believe robot designers should also be extremely careful about how they go about accomplishing these goals. While it is critical for *roboticists* to consider these factors and interrogate these power structures (see similar arguments regarding gender made by Winkle et al. [104]), I have argued that there are unjustifiable ethical risks to enabling *robots* to pursue these goals through means that would require computational representation, recognition, or reasoning related to these identity factors. There are a variety of ways that roboticists can choose to design robots with sensitivity to these risks.

One design path that roboticists could take moving forward would be to simply avoid giving robots anything approaching the ability to recognize, represent, or reason over identity characteristics in any way. Going a step farther, roboticists could even avoid capabilities that would seem to imply such capabilities and behaviors. For example, roboticists could avoid describing people in terms of physical descriptions altogether. As an example, in other, unpublished areas of our research, we have been working to evaluate robotic cognitive models of referring expression generation. To do so, we have been using simple "Guess Who" games in which participants and robots need to describe faces to one another and interpret each others' descriptions. This context embodies many of the nuances discussed in this paper.

Because we did not want our robots to need to give (or interpret) facial descriptions that used ostensible pigmentation or racial categorizations, we did not include such descriptors in our robot's knowledge base; and we selected a set of (cartoon) face stimuli which would not require the use of such descriptors for disambiguation. A consequence of this is that all of the faces used were likely to be racialized as white; and due to a lack of consideration of gender in our design, the problems we were trying to avoid for race reared their head for gender nonetheless, with robots and their interactants needing to rely on gendering in their descriptions. In future work, we are addressing this dilemma by simply moving away from face description. Nevertheless, we need to remain cognizant that the algorithms we design for other domains would, when reapplied to face domains by others, re-raise all of these concerns, paralleling the risks of clarification algorithms moved to ethically fraught contexts demonstrated by Jackson and Williams[43, 44].

A less drastic path forward, similar to image classification suggestions proposed by Scheuerman et al. [81] and foreshadowed by some of the discussion earlier in this section, would be to maintain robots' ability to perceive, describe and reason about descriptions of people, but to do so in ways that embrace ambiguity and center only what is observable, rather than making specific mention to categories. For example, Scheuerman et al. [81] suggests computer vision systems to classify people in terms of properties like whether they have a beard, whether they are wearing a dress, and so forth. These properties are important to peoples' identities and are key to what humans use to infer their own identity conclusions. Robots could similarly rely on other gender, race, and culturally *relevant* cues, without making conclusions about gender, culture, or race themselves.

Another way of framing these possible paths forward is through the lens of ethical robot design. Moor [63], 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 [83]. However, this same framework could similarly be used to argue for explicitly designing for *implicit ethical agency* in use contexts where concerns surrounding autonomous moral agents arise [94]. 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 identifying 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 [64] (see also [72]), but is not explicitly recognized, represented, or reasoned over by robots – or rather, by roboticists, through their robots.

7 CONCLUSION

In this paper, I briefly considered the myriad risks of attempting to recognize, represent, and/or reason about identity characteristics like race in interactive robotic systems. While roboticists should be cognizant of the dangers of colorblindness, and of the utility of robots with *implicit* sensitivity to race and other identity characteristics, I have argued that in most cases roboticists should refrain from designing robots that explicitly computationally recognize, represent, or reason over identity characteristics in their work. I hope that this paper brings awareness to these risks, and helps roboticists to steer away from dangerous racialized robotics technologies 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.

ACKNOWLEDGMENTS

I would like to thank the students in Spring 2022's "HNRS 498AA: De-Coding the Human Race" and "CSCI/HASS 432/532: Robot Ethics" for serving as a sounding board for my ideas while this paper was written.

This paper was funded in part by National Science Foundation awards IIS-1909847 and IIS-2044865, and by Air Force Young Investigator Award 19RT0497.

REFERENCES

[1] 2020. "No Justice, No Robots". https://nojusticenorobots.github.io/.
[2] Yasmeen Abu-Laban. 2015. Gendering surveillance studies: The empirical and normative promise of feminist methodology. *Surveillance & Society* 13, 1 (2015), 44–56.
[3] Michelle Alexander. 2012. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* The New Press.
[4] Abdullah Alzahrani, Simon Robinson, and Muneeb Ahmad. 2022. Exploring Factors Affecting User Trust Across Different Human-Robot Interaction Settings and Cultures. In *Human-Agent Interaction.* ACM.
[5] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of culture on the credibility of robot speech: A comparison between english and arabic. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction.* 157–164.
[6] Y Gavriel Ansara. 2012. Cisgenderism in medical settings: How collaborative partnerships can challenge structural violence. *Out of the ordinary: LGBT lives* 102 (2012), 1224.
[7] Anthony Appiah. 1994. Race, culture, identity: Misunderstood connections. (1994).
[8] Kirstie S Ball, David J Phillips, Nicola Green, and Hille Koskela. 2009. Surveillance studies needs gender and sexuality. *Surveillance & Society* 6, 4 (2009), 352–355.
[9] Christoph Bartneck, Toru Takahashi, and Yasuhiro Katagiri. 2004. Cross-cultural study of expressive avatars. In *Social Intelligence Design.*
[10] Christoph Bartneck, Kumar Yogeeswaran, Qi Min Ser, Graeme Woodward, Robert Sparrow, Siheng Wang, and Friederike Eyssel. 2018. Robots and racism. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction.* 196–204.
[11] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces* (2019).
[12] Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "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.* 1–19.
[13] Hector Betancourt and Steven R López. 1993. The study of culture, ethnicity, and race in American psychology. *American psychologist* 48, 6 (1993), 629.
[14] Eduardo Bonilla-Silva. 1997. Rethinking racism: Toward a structural interpretation. *American sociological review* (1997), 465–480.
[15] Eduardo Bonilla-Silva. 2004. From bi-racial to tri-racial: Towards a new system of racial stratification in the USA. *Ethnic and racial studies* 27, 6 (2004), 931–950.
[16] Eduardo Bonilla-Silva. 2006. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States.* Rowman & Littlefield Publishers.
[17] Simone Browne. 2010. Digital epidermalization: Race, identity and biometrics. *Critical Sociology* 36, 1 (2010), 131–150.
[18] Simone Browne. 2015. *Dark matters.* Duke University Press.
[19] Joy Adowaa Buolamwini. 2017. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers.* Ph. D. Dissertation. Massachusetts Institute of Technology.
[20] Stephen Cave and Kanta Dihal. 2020. The whiteness of AI. *Philosophy & Technology* 33, 4 (2020), 685–703.
[21] Michelle Christian. 2019. A global critical race and racism framework: Racial entanglements and deep and malleable whiteness. *Sociology of Race and Ethnicity* 5, 2 (2019), 169–185.
[22] Mark Coeckelbergh. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and information technology* 12, 3 (2010), 209–221.
[23] Beth Coleman. 2009. Race as technology. *Camera obscura: feminism, culture, and media studies* 24, 1 (2009), 177–207.
[24] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.
[25] Oliver C Cox. 1945. Race and caste: A distinction. *Amer. J. Sociology* 50, 5 (1945), 360–368.
[26] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, A Sánchez, et al. 2019. AI Now 2019 Report. New York: AI Now Institute.
[27] Will Culpepper, Thomas A. Bennett, Lixiao Zhu, Rafael Sousa Silva, Ryan Blake Jackson, and Tom Williams. 2022. IPOWER: Incremental, Probabilistic, Open-World Reference Resolution. In *Annual Meeting of the Cognitive Science Society (CogSci).*
[28] Rachel E Dubrofsky and Shoshana Amielle Magnet. 2015. *Feminist surveillance studies.* Duke University Press.
[29] Charles-Olivier Dufresne-Camaro, Fanny Chevalier, and Syed Ishtiaque Ahmed. 2020. Computer vision applications and their ethical risks in the global south. *Graphics Interface* (2020).
[30] Friederike Eyssel and Steve Loughnan. 2013. "It Don't Matter If You're Black or White"?. In *International Conference on Social Robotics.* Springer, 422–431.
[31] Karen Farquharson. 2007. 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).*
[32] Joe R Feagin. 2000. *Racist America: Roots, current realities, and future reparations.* Routledge.
[33] Susan Flynn and Antonia Mackay. 2018. *Surveillance, race, culture.* Springer.
[34] Pasquale Foggia, Antonio Greco, Gennaro Percannella, Mario Vento, and Vincenzo Vigilante. 2019. A system for gender recognition on mobile robots. In *Proceedings of the 2nd international conference on applications of intelligent systems.* 1–6.
[35] Marlena R Fraune, Yusaku Nishiwaki, Selma Sabanović, Eliot R Smith, and Michio Okada. 2017. Threatening flocks and mindful snowflakes: How group entitativity affects perceptions of robots. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.* 205–213.
[36] George M Fredrickson. 1981. *White supremacy: A comparative study of American and South African history.* OUP USA.
[37] Alessio Galatolo, Gaspar I Melsión, Iolanda Leite, and Katie Winkle. 2022. The Right (Wo) Man for the Job? Exploring the Role of Gender when Challenging Gender Stereotypes with a Social Robot. *International Journal of Social Robotics* (2022), 1–15.
[38] Clare Garvie. 2016. *The perpetual line-up: Unregulated police face recognition in America.* Georgetown Law, Center on Privacy & Technology.
[39] Norina Gasteiger, Mehdi Hellou, and Ho Seok Ahn. 2021. Factors for Personalization and Localization to Optimize Human–Robot Interaction: A Literature

Review. *International Journal of Social Robotics* (2021), 1–13.

[40] Jaap Ham. 2021. Influencing robot influence: Personalization of persuasive robots. *Interaction studies* 22, 3 (2021), 464–487.

[41] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems.* 1–13.

[42] Noel Ignatiev. 2012. *How the Irish became white.* Routledge.

[43] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 401–410.

[44] Ryan Blake Jackson and Tom Williams. 2022. Enabling morally sensitive robotic clarification requests. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 2 (2022), 1–18.

[45] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the role of gender in perceptions of robotic noncompliance. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction.* 559–567.

[46] Jessica K. Barfield. 2021. Discrimination and Stereotypical Responses to Robots as a Function of Robot Colorization. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* 109–114.

[47] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.

[48] Os Keyes, Josephine Hoy, and Margaret Drouhard. 2019. Human-computer insurrection: Notes on an anarchist HCI. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–13.

[49] Boyoung Kim and Joanna Korman. 2022. Treading lightly toward behavior change: Moral feedback from a robot on microaggressions. In *DEI Workshop at HRI 2022.*

[50] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 2021. Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* 10–18.

[51] Hille Koskela. 2012. "You shouldn't wear that body": The problematic of surveillance and gender. In *Routledge handbook of surveillance studies.* Routledge, 49–56.

[52] Matthew Le Bui and Safiya Umoja Noble. 2020. We're missing a moral framework of justice in artificial intelligence. In *The Oxford Handbook of Ethics of AI.* Oxford University Press, 163.

[53] Jon A Leydens and Juan C Lucena. 2017. *Engineering justice: Transforming engineering education and practice.* John Wiley & Sons.

[54] Dingjun Li, Pei-Luen Rau, and Ye Li. 2010. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics* 2, 2 (2010), 175–186.

[55] Jason Lim and Kath Browne. 2009. Senses of gender. *Sociological Research Online* 14, 1 (2009), 75–88.

[56] Velvetina Lim, Maki Rooksby, and Emily S Cross. 2021. Social robots on a global stage: establishing a role for culture during human–robot interaction. *International Journal of Social Robotics* 13, 6 (2021), 1307–1333.

[57] Timm Linder, Sven Wehner, and Kai O Arras. 2015. Real-time full-body human gender recognition in (RGB)-D data. In *2015 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 3039–3045.

[58] Michael L. Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfield, Michael C. Horowitz, Charles Isbell, Hiroaki Kitano, Karen Levy, Terah Lyons, Melanie Mitchell, Julie Shah, Steven Sloman, Shannon Vallor, , and Toby Walsh. 2021. Gathering Strength, Gathering Storms: The One Hundred YEar Study on Artificial Intelligence (AI100) 2021 Study Panel Report. http://ai100.stanford.edu/2021-report.

[59] Maxim Makatchev, Reid Simmons, Majd Sakr, and Micheline Ziadee. 2013. Expressing ethnicity through behaviors of a robot character. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 357–364.

[60] Serena Marchesi, Cecilia Roselli, and Agnieszka Wykowska. 2021. Cultural values, but not nationality, predict social inclusion of robots. In *International Conference on Social Robotics.* Springer, 48–57.

[61] Kevin A McLemore. 2015. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity* 14, 1 (2015), 51–74.

[62] Torin Monahan. 2009. Dreams of control at a distance: Gender, surveillance, and social control. *Cultural Studies? Critical Methodologies* 9, 2 (2009), 286–305.

[63] James H Moor. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21, 4 (2006), 18–21.

[64] Jaye Nias, Lelia Hampton, Princess Sampson, and Margie Ruffin. 2020. Decolonizing Technologies for Preserving Cultural and Societal Diversity. In *CHI 2020 Workshop on Engaging in Race in HCI.*

[65] Michael Omi and Howard Winant. 1993. On the theoretical status of the concept of race. *Race, identity and representation in education* (1993), 3–10.

[66] Patricia O'Neill-Brown. 1997. Setting the stage for the culturally adaptive agent. In *Proceedings of the 1997 AAAI fall symposium on socially intelligent agents.* AAAI Press Menlo Park, CA, 93–97.

[67] Pierre-Henri Orefice, Mehdi Ammi, Moustapha Hafez, and Adriana Tapus. 2016. Let's handshake and i'll know who you are: Gender and personality discrimination in human-human and human-robot handshaking interaction. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids).* IEEE, 958–965.

[68] Anastasia K Ostrowski, Raechel Walker, Madhurima Das, Maria Yang, Cynthia Breazea, Hae Won Park, and Aditi Verma. 2022. Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* IEEE, 969–976.

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

[70] Giulia Perugia and Dominika Lisy. 2022. Robot's Gendering Trouble: A Scoping Review of Gendering Humanoid Robots and its Effects on HRI. *arXiv preprint arXiv:2207.01130* (2022).

[71] Arnaud Ramey and Miguel A Salichs. 2014. Morphological gender recognition by a social robot and privacy concerns. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 272–273.

[72] Yolanda A Rankin and India Irish. 2020. A Seat at the Table: Black Feminist Thought as a Critical Framework for Inclusive Game Design. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[73] Laurel Riek and Don Howard. 2014. A code of ethics for the human-robot interaction profession. *Proceedings of we robot* (2014).

[74] Andrea J Ritchie. 2017. *Invisible no more: Police violence against Black women and women of color.* Beacon press.

[75] Selma Šabanović. 2010. Robots in society, society in robots. *International Journal of Social Robotics* 2, 4 (2010), 439–450.

[76] Alessia Saggese, Mario Vento, and Vincenzo Vigilante. 2019. MIVIABot: a cognitive robot for smart museum. In *International Conference on Computer Analysis of Images and Patterns.* Springer, 15–25.

[77] Anara Sandygulova, Mauro Dragone, and Gregory MP O'Hare. 2014. Real-time adaptive child-robot interaction: Age and gender determination of children based on 3d body metrics. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, 826–831.

[78] Elaheh Sanoubari, Stela H Seo, Diljot Garcha, James E Young, and Verónica Loureiro-Rodríguez. 2019. Good Robot Design or Machiavellian? An In-The-Wild Robot Leveraging Minimal Knowledge of Passersby's Culture. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 382–391.

[79] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[80] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8, 2 (2021), 20539517211053712.

[81] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[82] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. 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* 4, CSCW1 (2020), 1–35.

[83] Matthias Scheutz. 2017. The case for explicit ethical agents. *Ai Magazine* 38, 4 (2017), 57–64.

[84] Audrey Smedley. 1998. " Race" and the construction of human identity. *American Anthropologist* 100, 3 (1998), 690–702.

[85] Audrey Smedley. 2007. Antecedents of the racial worldview. *Race and racilization: Essential readings* (2007), 31–44.

[86] Audrey Smedley and Brian D Smedley. 2005. Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. *American psychologist* 60, 1 (2005), 16.

[87] Audrey Smedley and Brian D Smedley. 2012. *Race in North America: Origin and evolution of a worldview.* Westview Press.

[88] Robert Sparrow. 2019. Do robots have race?: Race, social construction, and HRI. *IEEE Robotics & Automation Magazine* 27, 3 (2019), 144–150.

[89] Robert Sparrow. 2020. Robotics has a race problem. *Science, Technology, & Human Values* 45, 3 (2020), 538–560.

[90] Megan Strait, Ana Sánchez Ramos, Virginia Contreras, and Noemi Garcia. 2018. Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 452–457.

[91] Thomas Trainer, John R Taylor, and Christopher J Stanton. 2020. Choosing the Best Robot for the Job: Affinity Bias in Human-Robot Interaction. In *International Conference on Social Robotics.* Springer, 490–501.

[92] Vilna Bashi Treitler. 2020. *The Ethnic Project.* Stanford University Press.

[93] Gabriele Trovato, Massimiliano Zecca, Salvatore Sessa, Lorenzo Jamone, Jaap Ham, Kenji Hashimoto, and Atsuo Takanishi. 2013. Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by Egyptians and Japanese. *Paladyn, Journal of Behavioral Robotics* 4, 2 (2013), 83–93.

[94] Aimee Van Wynsberghe and Scott Robbins. 2019. Critiquing the reasons for making artificial moral agents. *Science and engineering ethics* 25, 3 (2019), 719–735.

[95] Bram Vanderborght and Allison Okamura. 2020. United against racism and a call for action [Ethical, legal, and societal issues]. *IEEE Robotics & Automation Magazine* 27, 3 (2020), 10–11.

[96] Steven Vethman, Jildau Bouwman, Mark A. Neerincx, and Cor J. Veenman. 2022. Fairness in Human-Robot Interaction: Disparate Treatment for Good. In *DEI Workshop at HRI 2022.*

[97] Alex S Vitale. 2017. *The end of policing.* Verso Books.

[98] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. 2010. When in Rome: the role of culture & context in adherence to robot recommendations. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 359–366.

[99] Isabel Wilkerson. 2020. *Caste: The origins of our discontents.* Random House.

[100] Tom Williams and Matthias Scheutz. 2015. POWER: A Domain-Independent Algorithm for Probabilistic, Open-World Entity Resolution. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

[101] Tom Williams and Matthias Scheutz. 2016. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Thirtieth AAAI Conference on Artificial Intelligence.*

[102] Tom Williams and Matthias Scheutz. 2017. Referring expression generation under uncertainty: Algorithm and evaluation framework. In *Proceedings of the 10th International Conference on Natural Language Generation.* 75–84.

[103] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Bršćić, Iolanda Leite, and Tom Williams. 2022. Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 120–129.

[104] Katie Winkle, Erik Lagerstedt, Ilaria Torre, and Anna Offenwanger. 2022. 15 Years of (Who) man Robot Interaction: Reviewing the H in Human-Robot Interaction. *ACM Transactions on Human-Robot Interaction* (2022).

[105] Katie Winkle, Donald McMillan, Maria Arnelid, Madeline Balaam, Katherine Harrison, Ericka Johnson, and Iolanda Leite. 2023. Feminist Human-Robot Interaction: Disentangling Power, Principles and Practice for Better, More Ethical HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI).*

[106] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. 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.* 29–37.