Philosophy & Technology (2021) 34:1775–1779
https://doi.org/10.1007/s13347-021-00486-z

RESEARCH ARTICLE

Race and AI: the Diversity Dilemma

Stephen Cave[1] · Kanta Dihal[1]

Received: 26 September 2021 / Accepted: 27 September 2021 / Published online: 8 October 2021

Abstract

This commentary is a response to 'More than Skin Deep' by Shelley M. Park (Park, More than skin deep: A response to "The Whiteness of AI", Philosophy & Technology, 2021), and a development of our own 2020 paper 'The Whiteness of AI'. We aim to explain how representations of AI can be varied in one sense, whilst not being diverse. We argue that Whiteness's claim to universal humanity permits a broad range of roles to White humans and White-presenting machines, whilst assigning a much narrower range of stereotypical roles to people of colour. Because the attributes of AI in the popular imagination, such as intelligence, power and passing as human, are associated by the White racial frame with Whiteness, AI is cast predominantly as White. Following Sparrow (Science, Technology, & Human Values 45:538–560, 2020), we suggest this presents a dilemma for those creating or representing AI. We discuss three possible solutions: avoiding anthropomorphisation, explicitly critiquing racial role-typing, and representing powerful AI as non-White.

Keywords Artificial intelligence · Robots · Critical race studies · Racialisation · Anthropomorphism · Whiteness

In her commentary 'More than Skin Deep', Shelley M. Park makes three important points: first, that gender interplays with Whiteness in the construction, representation, marketing, and functionality of AI systems; second, that there are many different ways in which Whiteness is scripted in portrayals and instantiations of AI; and third, that White racial framing exceeds White casting and thus cannot be undone simply by more diverse and inclusive hiring in the mainstream culture industry (Park, 2021). We agree. However, we argue that the variety of portrayals of intelligent machines is nonetheless racially exclusive. That is, it is only machines presenting as White that are permitted a range of scripts; the roles permitted to those

✉ Stephen Cave
sjc53@cam.ac.uk

Kanta Dihal
ksd38@cam.ac.uk

[1] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

presenting as other than White are strictly limited. Building on Robert Sparrow, we argue that this poses a dilemma for AI engineers and roboticists who wish to address this lack of diversity, and examine three potential solutions.

1 Whiteness and Generality

Park notes that White AI systems, whether in fiction or as anthropomorphic robots, have a wide variety of attributes, from the cuteness of the Pepper robots to the menace of the Terminator. These vary according to a correspondingly wide range of scripts, shaped in particular (but not only) by gender stereotypes. According to Park, this variety show that not all White AI systems convey the kind of attributes that we argue are symptomatic of both Whiteness and AI, such as intelligence, professionalism and power. We agree that this variety exists. We argue, however, that whilst White machines take various forms, (a) these forms conform to clear themes that relate to these attributes of AI; and (b) machines racialised in other ways are much more limited in their roles and attributes.

Crucial to the function—and insidiousness—of the White racial frame is the idea that White people are the universal humans: that, in Western culture, they (alone) can be anything (Dyer, 1997, p. 2). Of course, within this frame, there are further distinctions along gender, class and ability lines. It would be more accurate to say that the middle- or upper-class able-bodied White male can be anything—the Man, in Sylvia Wynter's term, who has colonised the idea of the human (Wynter, 2003). Because White people are permitted to be anything, they do not always need to appear as heroic individuals or innocent maids; as universal humans, they can (and do) equally appear as evil geniuses or greedy thugs. When they appear in a negative light, it is not a vilification of their race: White people stand for themselves alone, the only people permitted to be individuals.

Not all roles played by White people therefore embody ideals of Whiteness; rather, roles that do embody these ideals can only be played by White people. It is not that all White people are portrayed as heroic individuals or innocent maids—but rather that non-White people (almost) never are. Thugs and thieves, on the other hand, can be of any colour. In contrast to a White person, a person of colour does represent their entire race, ethnicity, or community in every role they play, and consequently is limited to the stereotypes assigned to that group. In the White racial frame, people of colour only get to represent part of humanity, rather than the full scope of what a human may be (Weheliye, 2014, p. 3).

Relating this to AI, as Park describes, there are a range of ways in which AI is portrayed, and indeed a range of ways in which White AI is portrayed. They do not all exemplify the key attributes of AI to the same extent, and these portrayals are also often used to explore themes that have nothing to do with technology, such as motherhood or mortality. Nonetheless, there is a well-established mythology of AI with which Whiteness interplays. As noted in our previous papers (Cave, 2020; Cave & Dihal, 2020), this mythology associates AI not only with intelligence but also inventiveness, power and generality—the ability to become anything. These are all regarded by the White racial frame as distinctively White attributes (in particular, White male attributes). When machines with these attributes are built or portrayed, they are therefore racialised overwhelmingly as White, with other groups represented only in limited, stereotypical ways.

This is exemplified in the 2014 film *Ex Machina*, which depicts three female AIs in a blatant racial hierarchy. In his quest to create AI, engineer Nathan has produced a range of androids which culminate in Ava, played by a White actress (Alicia Vikander). Nathan introduces his guest Caleb to Ava, tasking Caleb to see if Ava can convince him of her humanity even though he can see that she is a machine—an audacious riff on the 'Turing test'. Ava is portrayed as intelligent, eloquent, creative and powerful—attributes the White racial frame associates with Whiteness. Indeed, her portrayal as a doe-eyed White beauty is intimately bound up with her presentation as fully human, even whilst she is a machine. As a consequence of these qualities, she passes the test, manipulating Caleb into letting her escape.

But a second, less obvious Turing test is being executed, which evidences the much more limited set of attributes associated with East Asian women in this film. Kyoko (Sonoya Mizuno) is a silent, assistive, feminine presence in Nathan's home, described by Nathan as unable to speak English. Caleb sees Kyoko being abused for spilling wine, and subsequently presenting herself as sexually available to both men. Only later does Kyoko reveal her artificial nature. She is an inferior iteration in Nathan's android project—unlike Ava unable to speak. Yet Caleb—and by implication, the viewer—does not question the idea that a real Asian woman would occupy such a diminished role, as we are all too familiar with the dehumanised stereotype of "the yellow woman, whose condition of objectification is often the very hope for any claims she might have to value or personhood" (Cheng, 2019, p. xi). The film's racial hierarchy is completed by the brief introduction of an even earlier iteration, the Black android Jasmine (Symara A. Templeman)—which is depicted without a head, not even able to pass as human in the limited way that Kyoko does. So, whilst three androids are presented, the White racial frame permits only one to be fully human-like AI.

2 The Diversity Dilemma

Park rightly argues that merely increasing the racial diversity of AI—real and fictional—will not address the harms created by its current predominant Whiteness if those AI systems are then portrayed in ways that merely perpetuate harmful stereotypes of those other racialised groups. We agree, as should be clear from our comments above. But as Robert Sparrow has noted, this presents roboticists, designers and artists with a dilemma. On the one hand, as we have noted, these machines have attributes of Whiteness and consequently have predominantly been portrayed as White. But at the same time, the role of these machines is frequently one of servant, or even slave, programmed to respond to the wants of their master. Given that the White racial frame has historically assigned people of colour to these positions, to racialise such machines as anything other than White could be seen as "reproducing and reinforcing traditional racist ideas about race and servitude" (Sparrow, 2020, p. 549).

We can imagine three possible solutions to this dilemma. First, as Sparrow advocates, designers could aim to avoid racialisation, or even anthropomorphism, altogether (Sparrow, 2020, p. 549). However, as he notes, a wide range of evidence from such fields as human-robotics interaction shows that anthropomorphism is frequently immensely helpful in establishing trust (or other desired relations) between human and machine. More specifically, there is evidence that *racialisation* can, in certain circumstances, be beneficial: for example, one recent study found that racial mirroring in chatbots (that is, allowing users to select avatars presenting the same racial identity as themselves) “had a positive influence on people’s perceived interpersonal closeness with the agent” and prompted “a higher desire to continue interacting with the agent” (Liao & He, 2020, p. 16). These positive effects were particularly pronounced for Black participants, suggesting racialisation can be important for building trust with marginalised communities.

Fortunately, two alternative solutions present themselves in popular culture, both of which aim to break down normative Whiteness and the associated limitations on roles for people of colour. On the one hand, several contemporary fictional depictions of AI explicitly critique the stereotyping of people—and AI—of colour in subjugated positions. Maeve in *Westworld* (Thandiwe Newton), for instance, breaks out of the role of whorehouse madam assigned to her by the White owners of the Westworld theme park, and leads the liberation of the park’s androids.

On the other hand, the Whiteness of AI can be disrupted simply by portraying as people of colour those machines that have previously been solely the domain of Whiteness. So, not the servants and sex slaves, but AI that is powerful, intelligent or benevolent, breaking the stereotypes of the White racial frame and the history of AI narratives together. A recent example is the smart, morally complex android soldier played by Black actor Anthony Mackie in Netflix’s 2021 film *Outside the Wire*.[1]

3 Moral Purity and the White Utopia

Finally, we have previously argued that the pervasive Whiteness of AI follows well-established visions of a White techno-Utopia in which people of colour have been replaced—even as slaves and servants—by White-presenting machines. Park rightly points out that, whilst this might allow White people to *feel* less guilty about their exploitation of others, in reality they remain guilty, as people of colour continue to be exploited to make these machines—just offshore and out of sight. Addressing this will require wholly different ways of framing AI—not as autonomous, humanoid entities created by a lone genius as Hollywood so often has it, but as constructs of

[1] *Contra* Park, we consider this a contrast with the Whiteness of the killer robots in the *Terminator* franchise. Whilst she argues that the exaggerated musculature of Schwarzenegger represents ‘dark’ savagery, we follow Dyer in the view that the perfectly chiselled muscular frame represents the conquest of mind over body—a capacity claimed by Whiteness (Dyer, 1997). In addition, we contend that the various Terminator models can only fulfil their missions by passing as White, with the privileges and access that come with that.

resources, labour and data. Alex Rivera's 2008 film *Sleep Dealer* is a rare example: portraying a near future in which Mexican workers perform essential labour in the USA by remotely operating robots from their hometowns. In this film, American citizens see the machines and enjoy the products of their labour, but do not see the exploited workforce that sustains them. Addressing AI's diversity dilemma is therefore only scratching the surface: fully addressing the Whiteness of AI will require more such works that reveal the inequalities and injustices hidden under the white plastic.

Acknowledgements The authors would like to thank Kerry Mackereth and Eleanor Drage for comments.

Funding Stephen Cave and Kanta Dihal are funded by the Leverhulme Trust (via grant number RC-2015–067 to the Leverhulme Centre for the Future of Intelligence). Kanta Dihal is additionally funded through the support of grants from DeepMind Ethics & Society and Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Cave, S., 2020. The problem with intelligence: Its value-laden history and the future of AI, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20. ACM, New York, pp. 29–35. https://doi.org/10.1145/3375627.3375813

Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00415-6

Cheng, A. A. (2019). *Ornamentalism*. Oxford University Press.

Dyer, R. (1997). *White*. Routledge.

Liao, Y., He, J., 2020. The racial mirroring effects on human-agent in psychotherapeutic conversation. *In Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI'20*. ACM, New York, pp. 430–442. https://doi.org/10.1145/1234567890

Park, S. M. (2021). *More than skin deep: A response to "The Whiteness of AI."* Philosophy & Technology.

Sparrow, R. (2020). Robotics has a race problem. *Science, Technology, & Human Values, 45*, 538–560. https://doi.org/10.1177/0162243919862862

Weheliye, A. G. (2014). *Habeas viscus: Racializing assemblages, biopolitics, and black feminist theories of the human*. Duke University Press.

Wynter, S. (2003). Unsettling the coloniality of being/power/truth/freedom: Towards the human, after man, its overrepresentation—An argument CR. *The New Centennial Review, 3*, 257–337.