Philosophy & Technology (2020) 33:685–703
https://doi.org/10.1007/s13347-020-00415-6

RESEARCH ARTICLE

The Whiteness of AI

Stephen Cave[1] · Kanta Dihal[1]

Received: 3 January 2020 / Accepted: 28 June 2020 / Published online: 6 August 2020

Abstract

This paper focuses on the fact that AI is predominantly portrayed as white—in colour, ethnicity, or both. We first illustrate the prevalent Whiteness of real and imagined intelligent machines in four categories: humanoid robots, chatbots and virtual assistants, stock images of AI, and portrayals of AI in film and television. We then offer three interpretations of the Whiteness of AI, drawing on critical race theory, particularly the idea of the White racial frame. First, we examine the extent to which this Whiteness might simply reflect the predominantly White milieus from which these artefacts arise. Second, we argue that to imagine machines that are intelligent, professional, or powerful is to imagine White machines because the White racial frame ascribes these attributes predominantly to White people. Third, we argue that AI racialised as White allows for a full erasure of people of colour from the White utopian imaginary. Finally, we examine potential consequences of the racialisation of AI, arguing it could exacerbate bias and misdirect concern.

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

> Overall, I construe race, racialization, and racial identities as on-going sets of political relations that require, through constant perpetuation via institutions, discourses, practices, desires, infrastructures, languages, technologies, sciences, economies, dreams, and cultural artefacts, the barring of nonwhite subjects from the category of the human as it is performed in the modern west.
> Alexander G. Weheliye (Weheliye 2014, 2)

> Technology as an abstract concept functions as a white mythology.
> Joel Dinerstein (Dinerstein 2006, 570)

✉ Stephen Cave
sjc53@cam.ac.uk

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

1 Introduction

It is a truth little acknowledged that a machine in possession of intelligence must be white. Typing terms like "robot" or "artificial intelligence" into a search engine will yield a preponderance of stock images of white plastic humanoids. Perhaps more notable still, these machines are not only white in colour, but the more human they are made to look, the more their features are made ethnically White.[1] In this paper, we problematize the often unnoticed and unremarked-upon fact that intelligent machines are predominantly conceived and portrayed as White. We argue that this Whiteness both illuminates particularities of what (Anglophone Western) society hopes for and fears from these machines, and situates these affects within long-standing ideological structures that relate race and technology.

Race and technology are two of the most powerful and important categories for understanding the world as it has developed since at least the early modern period. Yet, as a number of scholars have noted, their profound entanglement remains understudied (Sinclair 2004; de la Peña 2010). There are a number of possible reasons for this—and, as Bruce Sinclair writes, "racial prejudice dominates all of them" (Sinclair 2004, 1). They include the lack of first- or secondhand accounts of the role of people of colour in the development and use of technology; persistent stereotypes about technology as the province and product of one particular racial group—White people; and the persistent tendency of members of that group, who dominate the academy in the US and Europe, to refuse to see themselves as racialised or race as a matter of concern at all.

This lack of scholarly attention is surprising because, as Michael Adas elucidated in 1989, the idea of technological superiority was essential to the logic of colonialism. Not only was superior weaponry and transportation (etc.) necessary for large-scale conquest and control of foreign territory, it was also part of its justification: proof that White Europeans were an advanced civilisation with a right to rule over others (Adas 1989). Fortunately, this lack of attention is increasingly being remedied, and the relationship between race and technology is beginning to garner the kind of attention that has since the 1970s been given to gender and technology, following the pioneering work of Donna Haraway, Sandra Harding, and Evelyn Fox Keller (Haraway 1991; Harding 1986; Keller 1985). This includes attention to this century's ubiquitous digital technologies. In 2006, Lisa Nakamura asked, "How do we make cyberculture studies a field that *as a matter of course* employs critical race theory and theories of cultural difference…?" (Nakamura 2006, 35). Since then, a number of significant works have attempted to do just that, including Safiya Noble's *Algorithms of Oppression* and Ruha Benjamin's *Race After Technology* (Noble 2018; Benjamin 2019).

This paper aims to contribute to this body of literature on race and technology by examining how the ideology of race shapes conceptions and portrayals of artificial intelligence (AI). Our approach is grounded in the philosophy of race and critical race theory, particularly the Black feminist theories of bell hooks, Sylvia Wynter and Alexander G. Weheliye (hooks 1992 1997; Wynter 2003; Weheliye 2014), and work

[1] Following the increasingly common usage of the capitalised form "Black" to denote the ethnicity and "black" the colour, we use "White" to refer to the ethnicity and "white" the colour. While not yet the norm, as can be seen in our quotations of critics who do not employ this distinction, this usage will make our discussion clearer.

in Whiteness studies, including that of Richard Dyer, Joe R. Feagin, and Ruth Frankenberg (Dyer 1997; Feagin 2013; Frankenberg 1997a). In 2006, Feagin coined the term "white racial frame" to describe those aspects of the Anglophone Western worldview that perpetuate a racialised hierarchy of power and privilege (Feagin 2006). In his words, "the white racial frame includes a broad and persisting set of racial stereotypes, prejudices, ideologies, interlinked interpretations and narratives, and visual images" (Feagin 2013, xi). Although it reached its peak in the age of colonial expansion, this framing persists: "Today, as whites move through their lives, they frequently combine racial stereotypes and biases (a beliefs aspect), racial metaphors and concepts (a deeper cognitive aspect), racialised images (the visual aspect), racialised emotions (feelings), interpretive racial narratives, and inclinations to discriminate within a broad racial framing" (Feagin 2013, 91). In essence, this paper examines how representations of AI reflect this White racial frame.

One of the main aims of critical race theory in general, and Whiteness studies in particular, is to draw attention to the operation of Whiteness in Western culture. The power of Whiteness's signs and symbols lies to a large extent in their going unnoticed and unquestioned, concealed by the myth of colour-blindness. As scholars such as Jessie Daniels and Safiya Noble have noted, this myth of colour-blindness is particularly prevalent in Silicon Valley and surrounding tech culture, where it serves to inhibit serious interrogation of racial framing (Daniels 2013, 2015; Noble 2018). Hence the first step for such an interrogation is, in Richard Dyer's term, to "make strange" this Whiteness, de-normalising and drawing attention to it (Dyer 1997, 10). As Steve Garner puts it, the reason "for deploying whiteness as a lens is that it strips a normative privileged identity of its cloak of invisibility" (Garner 2007, 5). This is our primary intention in examining intelligent machines through the White racial frame.

In the next section of this paper, we first lay out current evidence for the assertion that conceptions and portrayals of AI—both embodied as robots and disembodied—are racialised, then evidence that such machines are predominantly racialised as White. In the third section of the paper, we offer our readings of this Whiteness. Our methods are qualitative. As de la Peña writes: "Studying whiteness means working with evidence more interpretive than tangible; it requires imaginative analyses of language and satisfaction with identifying possible motivations of subjects, rather than definitive trajectories of innovation, production, and consumption" (de la Peña 2010, 926). We offer three interpretations of the Whiteness of AI. First, the normalisation of Whiteness in the Anglophone West can go some way to explaining why that sphere's products, including representations of AI, are White. But we argue that this argument alone is insufficient. Second, we argue that to imagine an intelligent (autonomous, agential, powerful) machine is to imagine a White machine because the White racial frame ascribes these attributes predominantly to White people. Thirdly, we argue that AI racialised as White allows for a full erasure of people of colour from the White utopian imaginary. Such machines are conceived as tools that will replace "dirty, dull, or dangerous" tasks (Murphy 2000, 16), including replacing human interactions that are considered metaphorically dirty: White robot servants will allow the White master to live a life of ease unsullied by interaction with people of other races.

2 Seeing the Whiteness of AI

Our concern in this paper is with the racialisation (as White) of both real and imagined machines that are implied or claimed to be intelligent. By racialisation, we mean the ascription of characteristics that are used to identify and delineate races in a given racial frame, which in this case is the Anglophone West. Feagan notes:

> Among the most important ingredients of this frame are: (1) the recurring use of certain physical characteristics, such as skin colour and facial features, to differentiate social groups; (2) the constant linking of physical characteristics to cultural characteristics; and (3) the regular use of physical and linked cultural distinctions to differentiate socially "superior" and "inferior" groups in a social hierarchy (Feagin 2013, 41).

It is worth noting that "physical characteristics" need not only refer to those that are visible: voice and accent are also used as markers for social categorisation. Similarly, the category "cultural characteristics" is also used expansively and can include markers such as dialect, mannerisms, and dress codes, as well as mental and moral qualities, such as diligence, industriousness, reliability, trustworthiness, inventiveness, and intellectual ability. Indeed, these mental and moral qualities have always been an essential part of the racial frame, as it is largely on the basis of these that claims of superiority or inferiority have been made.

2.1 Machines Can Be Racialised

That machines can be racialised, in the sense that they can be given attributes that enable their identification with human racial categories, has been empirically demonstrated. For example, in one study, Christoph Bartneck and colleagues took pictures of the humanoid Nao robot and adjusted the colouration to match the skin tone of stock images of White and Black people (Bartneck et al. 2018). They then asked participants to define the race of the robot with several options including "does not apply". A minority—ranging across the experiments from 7 to 20%—chose the "does not apply" option, while a majority—ranging from 53 to 70%—identified the robots as belonging to the race from which their colouration derived. They concluded "Participants were able to easily and confidently identify the race of robots according to their racialization [...] Thus, there is also a clear sense in which these robots – and by extension other humanoid robots – do have race" (Bartneck et al. 2018, 201).

This should not be surprising. Many machines are anthropomorphised—that is, made to be human-like to some degree—in order to facilitate human-machine interaction. This might involve obvious physical features (a head on top, two eyes, a mouth, four limbs, bipedalism, etc.), but it can also include invisible features such as a human-like voice, or human-like interactions, such as politeness or humour. Given the prevalence of racial framing, in most contexts, to be human-like means to have race. Consequently, as Liao and He point out in their discussion of the racialisation of psychotherapeutic chatbots, "racial identity is an integral part of anthropomorphized agents" (Liao and He 2020, 2). They go on to explore a number of racial cues for virtual agents, including visual cues such as skin colour, but also cultural signifiers such as

names (e.g. for male names, Jake as White, Darnell as Black, and Antonio as Hispanic). Similarly, "even text-based conversational exchanges"—that is, those with no visual component at all—"perform a racial or ethnic identity" through the interlocutors' choice of dialect, etc. (Marino 2014, 3).

Given the sociopolitical importance of the racial frame in structuring people's interactions, if machines are really being racialised, then we would expect this to have an impact on how people interact with these machines. Numerous studies show just this. For example, Liao and He found that a person's "perceived interpersonal closeness" with a virtual agent is higher when the virtual agent has the same racial identity as that person (Liao and He 2020, 2). Other studies reflect the extent to which racism—prejudicial treatment on the basis of race—is intrinsic to racial framing.

As detailed in their paper "Robots Racialized in the Likeness of Marginalized Social Identities are Subject to Greater Dehumanization than Those Racialized as White", Strait et al. analysed free-form online responses to three videos, each depicting a female-gendered android with a different racial identity: Black, White, and East Asian. Their aim was to assess whether the same kind of marginalising and dehumanising commentary that is applied to real people of colour would be applied to these robots. They found that the valence of the commentary was significantly more negative towards the Black robot than towards the White or Asian ones and that both the Asian and Black robots were subject to over twice as many dehumanising comments as the White robot (Strait et al. 2018).

Two recent studies have further examined the transfer of bias to machines using the "shooter bias" paradigm. This paradigm was first described in the 2002 paper "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals" (Correll et al. 2002). It used a simple video game featuring images of (real) Black and White male targets, each holding either a gun or a nonthreatening object. Participants were instructed to shoot only armed targets. A clear racial bias was identified: "participants fired on an armed target more quickly when he was African American than when he was White, and decided not to shoot an unarmed target more quickly when he was White than when he was African American" (Correll et al. 2002, 1325). Studies by Bartneck et al. and Addison et al. used the same methodology to examine whether this "shooter bias" would be transferred to racialised robots (Bartneck et al. 2018; Addison et al. 2019). They found that "people showed a similar shooter bias toward robots racialized as Black relative to White in a similar fashion as they showed toward Black vs. White humans, no matter their own race" (Addison et al. 2019, 493).

2.2 Whiteness as the Norm for Intelligent Machines

The previous section shows that research has empirically demonstrated that machines can be racialised and that this racialisation includes transfer of the attendant biases found in the human world. In this subsection, we will survey evidence for the extent to which AI systems—machines purported to be intelligent—are predominantly racialised as White. We will look briefly at four categories: real humanoid robots, virtual personal assistants, stock images of AI, and portrayals of AI in film and television.

2.2.1 The Whiteness of Humanoid Robots

A number of commentators have remarked on the preponderant Whiteness of humanoid robots. In their proposed "code of ethics" for human-robot interaction Riek and Howard note the "lack of diversity in robot morphology and behavior":

> In terms of race, with precious few exceptions, such as Hanson's Bina48, the vast majority of android and gynoid robots are Asian or Caucasian in their features for no discernible reason. Furthermore, most of these robots tend to have a euro-centric design with regards to their appearance, behavior, and voice. (Riek and Howard 2014, 4)

Human-computer interaction researchers Christoph Bartneck and colleagues, who conducted some of the studies cited above, have also noted that robots are usually racialised as White: "most of the main research platforms for social robotics, including Nao, Pepper, and PR2, are stylized with white materials and are presumably White" (Bartneck et al. 2018, 202). Finally, media studies and literary scholar Jennifer Rhee notes the "normalization and universalization of whiteness" as expressed both in earlier robotics research and in robot toys: "Kismet, with its blue eyes, light brown eyebrows, and pink ears, also 'normalizes whiteness', as do other robot companions, such as the blonde-haired, blue-eyed Cindy Smart Doll and the similarly blonde-haired, blue-eyed My Friend Cayla." (Rhee 2018, 105).

Although robots such as Nao and Pepper have enjoyed commercial success, neither has received quite the attention garnered by Sophia from Hanson Robotics. This machine consists foremost of a White humanoid head, sometimes also with an upper torso (see Fig. 1). It has not only given numerous high-profile television interviews but also received political honours, including in 2017 receiving citizenship of Saudi Arabia and becoming an "Innovation Champion" for the United Nations Development Programme (Weller 2017; UNDP 2017).

2.2.2 The Whiteness of Chatbots and Virtual Assistants

Though conversational agents do not exhibit any visual racial cues, they are racialised by means of sociolinguistic markers (Sweeney 2016; Villa-Nicholas and Sweeney 2019). Discussing ELIZA, an influential natural language processing program created by Joseph Weizenbaum at the MIT AI Laboratory in 1966, Mark Marino writes: "If ELIZA presented a bot that tried to imitate language, it was performing standard white middle-class English, without a specific identifying cultural inflection... language without culture, disembodied, hegemonic, and, in a word, white" (Marino 2014, 5). Since then, natural language processing has entered the mainstream, with "virtual assistants" existing in many people's pockets, handbags, or homes through devices such as smartphones. Indeed, this is one of the most common ways in which people interact with technology that could be labelled "AI". These tools present their designers with many decisions about socio-cultural positioning. Ruha Benjamin recalls this anecdote:

> A former Apple employee who noted that he was "not Black or Hispanic" described his experience on a team that was developing speech recognition for

Fig. 1 Sophia. Hanson Robotics, April 2020

> Siri, the virtual assistant program. As they worked on different English dialects — Australian, Singaporean and Indian English — he asked his boss: "What about African American English?" To this his boss responded: "Well, Apple products are for the premium market." (Benjamin 2019, 28)

As a further example, she describes a Black computer scientist who chose a White voice for his app rather than a Black one, so as not to "create friction" (Benjamin 2019, 28–29). So while some designers might be unconsciously racialising their products as White, others are doing so in full awareness of this choice.

2.2.3 The Whiteness of Stock Images of AI

As anyone working in the field will know, stock images of AI, at least when anthropomorphised, are overwhelmingly white and arguably overwhelmingly White. The more realistically humanoid these machines become, the more Caucasian in their features. Such images are used to illustrate not only generalist newspaper articles and corporate slideshows but also specialist and technical works, and even works of a critical nature, such as Harry Collins's *Artifictional Intelligence* (Polity, 2018) and Anthony Elliott's *The Culture of AI* (Routledge, 2018) (Fig. 2).

The prevalence of such images is reflected in the results of search engines. Such searches are a useful indicator of how a subject is portrayed at a given time, for two reasons. First, search engines are very widely used (approximately 3.5 billion searches are made on Google every day, or 40 thousand per second[2]) and can therefore be considered a highly influential source of information and perceptions. Second, the nature of such search engines means that they are not only promoting certain ideas and perceptions but also reflecting their existing prevalence. While the exact nature of

[2] https://www.internetlivestats.com/google-search-statistics/ accessed 30 December 2019.

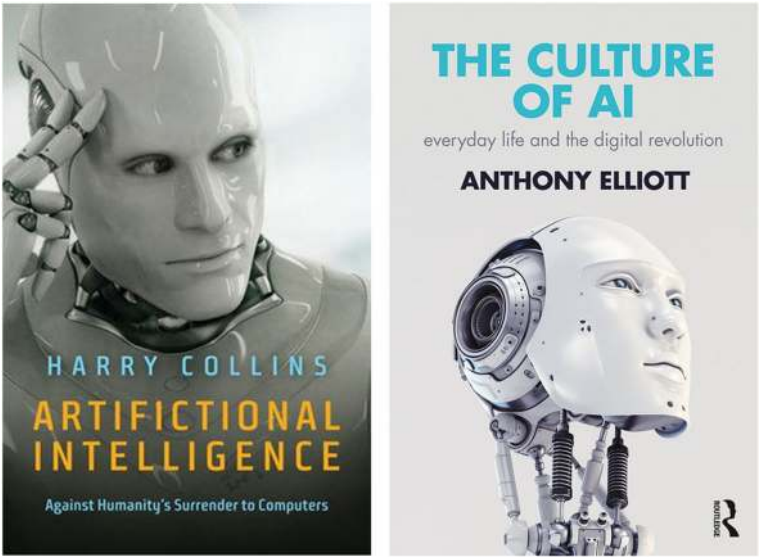

Fig. 2 Covers of Collins 2018, Polity, and Elliott 2018, Routledge

Google's search, ranking, and result presentation algorithms is proprietary, we know that they evaluate (crudely put) influence and popularity—for example, in terms of how many other sites link to a given website. So the fact that certain images are shown when someone searches for a relevant term means not only that those images are being thus promoted by some of the most powerful organs of content mediation in existence today but also that these images are already widespread and used on other influential websites, as that is what underlies their promotion by the search engines.

Consequently, search results are increasingly examined by scholars, including in the study of racial bias. For example, in her 2018 book *Algorithms of Oppression: How Search Engines Reinforce Racism*, Safiya U. Noble identifies many ways in which such sites reflect and exacerbate prejudice, such as the search results for "Latinas" that feature mostly porn (Noble 2018, 75, 155) or the White men who come up when searching for images of professions such as "construction worker", "doctor", or "scientist" (Noble 2018, 82–83).

In order to get an indication of the prevalence of these racialised machines on the internet, we conducted two image searches on Google (the most widely used search engine) using the anonymous Tor browser to ensure results were not influenced by our personal search histories and locations. We first searched on the term "artificial intelligence": the top results are in Fig. 3. Some of these results are too abstract, featuring stylised brains and circuits, for example, to be considered racialised. However, among the results showing humanoid figures, racialisation as White predominates. First, two pictures show actual human hands, and both are White. Second, a further two pictures show humanoid robots, and both are White and could thus be read as White, as Bartneck et al. suggest (Bartneck et al. 2018, 202). Therefore, we might say that inasmuch as the machines are racialised, they are racialised as White.

In order to focus more on representations of embodied, anthropomorphic AI, we also searched for "artificial intelligence robot": the top results are in Fig. 4. As is clear, this

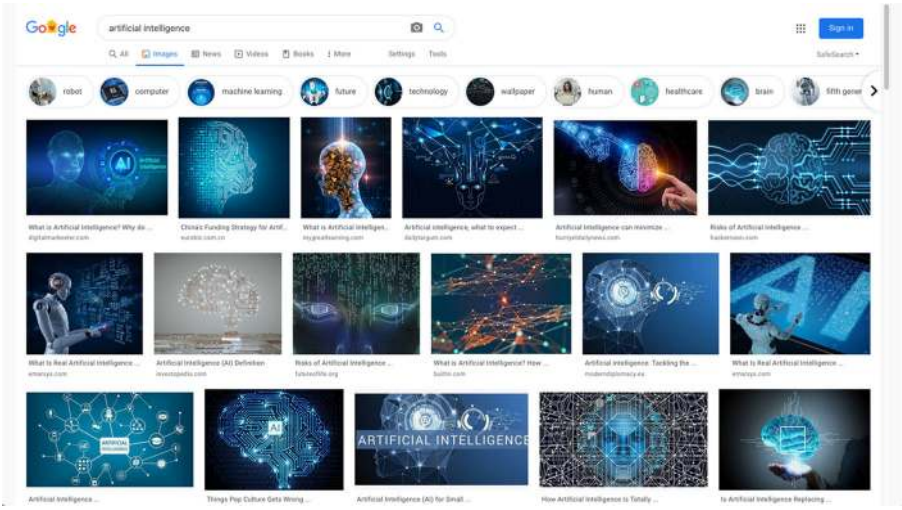

Fig. 3 Tor browser Google image search result for "artificial intelligence", 13 April 2020

search produces an even greater preponderance of images that are either white in colour or racialised as White or both.

2.2.4 The Whiteness of AI in Film and Television

These contemporary stock images distil the visualisations of intelligent machines in Western popular culture as it has developed over decades. In science fiction from the nineteenth century onwards, AI is predominantly imagined as White. For example, the Terminator (Arnold Schwarzenegger), RoboCop (Peter Weller and Joel Kinnaman), all of the "replicants" in the *Blade Runner* franchise (e.g. Rutger Hauer, Sean Young, and Mackenzie Davis), Sonny in *I, Robot* (Alan Tudyk), Ava in *Ex Machina* (Alicia

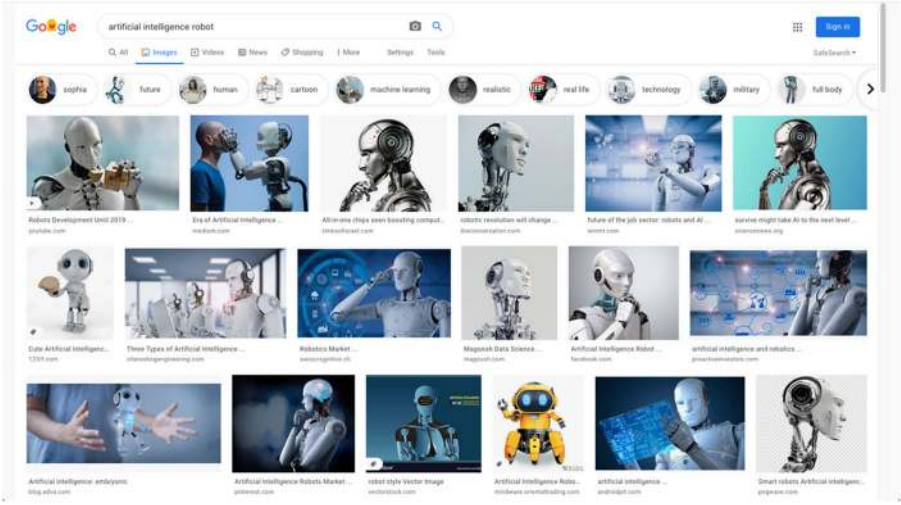

Fig. 4 Tor browser Google image search result for "artificial intelligence robot", 13 April 2020

Vikander) (Fig. 5), and Maria in *Metropolis* (Brigitte Helm) are all played by White actors and are visibly White on screen. Androids made of metal or plastic are also usually given White facial features, such as the robots in the 2007 film *I, Robot*.

Even disembodied AI is imagined as White: HAL-9000 in *2001: A Space Odyssey* and Samantha in *Her* are voiced by White actors. All of these AIs come from Hollywood films; they have been produced in a country in which 18% of the population is Hispanic, but in which only one fictional robot has that background: Bender Rodríguez in the animated TV series *Futurama*, who is canonically constructed in Mexico—but who is voiced by the White voice actor John DiMaggio. Only very recent TV shows with a large cast of androids, such as *Westworld* and *Humans*, have attempted to address this with AI characters evincing a mix of skin tones and ethnicities. This preponderance of intelligent machines racialised as White led Dyer to posit "the android as a definition of whiteness" (Dyer 1997, 213).

3 Understanding the Whiteness of AI

We offer three interpretations of the racialisation of intelligent machines as White: the Whiteness of their creators perpetuating itself; the Whiteness of the attributes ascribed to AI; and the extent to which AI permits the erasure of people of colour from the White utopia.

3.1 Whiteness Reproducing Whiteness

In European and North American societies, Whiteness is normalised to an extent that renders it largely invisible. As Toby Ganley puts it in his survey of Whiteness studies, "the monopoly that whiteness has over the norm" is one of the field's two unifying insights—the other being that it confers power and privilege (Ganley 2003, 12). Richard Dyer describes this as the view that "other people are raced, we are just people" (Dyer 1997, 1). This normalisation means that Whiteness is not perceived by majority populations as a distinct colour, but rather as an absence of colour—colour

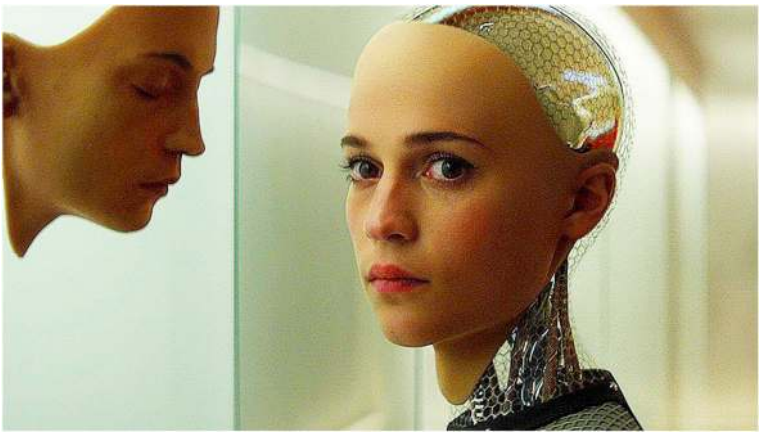

Fig. 5 Alicia Vikander as Ava in *Ex Machina*. Source: Youtube

both in the literal sense and in the sense of race. Consequently, the Whiteness of AI could be considered simply a default. It does not appear as a feature, but is transparent, like the air we breathe: the "unmarked marker", as Ruth Frankenberg calls it (Frankenberg 1997b, 1). The majority of White viewers are unlikely to see human-like machines as racialised at all, but simply as conforming to their idea of what "human-like" means.

For non-White people, on the other hand, Whiteness is never invisible in this manner, as bell hooks reminds us (hooks 1992 1997). So-called colour-blindness, an attitude of not seeing race, and of presuming that people in contemporary society are no longer disadvantaged on the basis of race, is itself a narrative that perpetuates White hegemony: "communities of color frequently see and name whiteness clearly and critically, in periods when white folks have asserted their own 'color blindness'" (Frankenberg 1997b, 4). Noble argues that "central to these 'colorblind' ideologies is a focus on the inappropriateness of 'seeing race'"—a view that she argues is dominant among Silicon Valley technologists, who "revel in their embrace of colorblindness as if it is an asset and not a proven liability" (Noble 2018, 168). Such colour-blindness is a liability because it obscures the normalisation of Whiteness and marginalisation of other racialised groups—and the real world effects this has, such as facial recognition technologies not distinguishing Black or East Asian faces (Buolamwini and Gebru 2018).

Given the normalisation of Whiteness, for some designers, to make a human-like machine will unthinkingly mean to make a White machine. As Dyer puts it: "white people create the dominant images of the world and don't quite see that they thus create the dominant images of the world in their own image" (Dyer 1997, 9). But this alone is not a satisfactory explanation of the Whiteness of AI, as not all entities—more specifically, not all intelligent, humanoid entities—imagined by predominantly White industries are portrayed as White. For example, Western science fiction has a long tradition of White authors racialising extraterrestrials as non-White. In the late nineteenth century, for instance, the real-world fear of the "Yellow Peril" was metaphorically addressed in science fiction by racialising extraterrestrial invaders as East Asian. The Flash Gordon franchise gained its lead villain in a 1934 comic, which introduced the tyrannical emperor of the planet Mongo—the Orientalised alien Ming the Merciless.

> Such is the villain in *Flash Gordon* - a trident bearded, slanty eyed, shiny doomed [sic], pointy nailed, arching eyebrowed, exotically garbed Oriental named Ming, who personifies unadulterated evil. A heavy like Ming is not contrived in a comic strip writer's imagination during a coffee break, but rather is the product of perhaps the richest and longest tradition of all of Hollywood ethnic stereotypes. (Barshay 1974, 24–26)

Dyer points out that *Blade Runner* similarly deliberately uses East Asian characters in order to offset the whiteness of its protagonists, including the White androids: "the yellow human background emphasises the chief protagonists' whiteness. The whitest of hue are the replicants" (Dyer 1997, 214). Racial stereotyping of aliens is not a phenomenon limited to past centuries. The *Star Wars* prequel trilogy (Lucas 1999; 2002; 2005) has been criticised for the "transparent racism" in its depiction of the alien

Jar Jar Binks as a West Indian caricature (Lavender 2011, 193) reminiscent of blackface minstrelsy (Williams 1999), and of the slave trader Watto, an antisemitic Jewish caricature with a large nose, skullcap, Yiddish accent, and obsession with money (Freedman 2019).

This racialisation of aliens in SF suggests that the racialisation of artificial intelligence is a choice. The White racial frame as perpetuated by the White creators of these works portrays dangerous invaders from another planet as East Asian and bumbling alien petty-criminals as Afro-Caribbean. Therefore, the fact that it portrays AI as overwhelmingly White requires further explanation. In the following sections, we offer two.

3.2 AI and the Attributes of Whiteness

While Whiteness functions in part through its invisibility in mainstream discourse, this does not mean it has no distinguishable features of its own. Indeed, the White racial frame has a long history of ascribing certain attributes to Whites and disputing them in others: these are the very claims that have been used to justify colonialism, segregation, and other modes of oppression. We argue that AI is predominantly racialised as White because it is deemed to possess attributes that this frame imputes to White people. We examine these attributes under three key headings: intelligence, professionalism, and power.

First, the primary attribute being projected onto these machines is, as the term "AI" suggests, intelligence. Throughout the history of Western thought, but in particular since the seventeenth century in Europe and the territories it colonised, intelligence has been associated with some humans more than others (Carson 2006). The idea that some races were more mentally able than others was crucial to the legitimation of the advancing colonial project. Those deemed less intelligent—in the words of Rudyard Kipling, "Half-devil and half-child"—were judged unqualified to rule themselves and their lands. It was therefore legitimate—even a duty, "the white man's burden" as Kipling put it—to destroy their cultures and take their territories (Kipling 1899). Through the nineteenth century, strenuous efforts were made to empirically demonstrate and measure this intellectual difference, culminating in the development of the IQ test (Gould 1981). Although explicit associations between racial groups and intelligence declined after the Second World War, (a) they continue to be made in right-wing circles (Saini 2019) and (b) implicit or unconscious associations between race and intelligence persist widely (see, for example, van den Bergh et al. 2010; Okeke et al. 2009). Given the White racial frame has for centuries promoted the association of intelligence with the White, European race, it is to be expected that when this culture is asked to imagine an *intelligent* machine, it imagines a White machine.

A crucial aspect of the idea of intelligence is generality. Intelligence is often defined as a "general mental capability" (Gottfredson 1997), and in AI, the concept of "artificial *general* intelligence"—a system with the kind of flexible mental capabilities humans have—is often considered to be the original and primary goal of the field (Crevier 1993). But in the White racial frame, not all humans are considered to have this attribute to the same degree. As Weheliye puts it, using Sylvia Wynter's idea of "the Man"—the Enlightenment, Western, White male subject, "In the context of the secular human, black subjects, along with indigenous populations, the colonised, the insane,

the poor, the disabled, and so on serve as limit cases by which Man can demarcate himself as the universal human" (Weheliye 2014, 24). According to the White racial frame, it is the rational, scientific thought of the White Westerner that lays claim to universal validity—or, we might say, true generality. Other races, by contrast, are framed as particular and subjective, constrained by the limits of their non-ideal bodies and cultures to think thoughts that are partial and parochial. To imagine a truly intelligent machine, one with general intelligence is therefore to imagine a White machine.

Second, much of the current discourse around AI focuses on how it is, or will soon be, capable of *professional* work. This is frequently claimed to be what makes the present wave of automation different from previous waves, in which machines became capable of supplanting manual and semi-skilled labour (Ford 2015). Professional work—law, medicine, business, and so forth—is at the upper end of pay and status scales. White Europeans and North Americans have historically not considered all humans equally fit for such roles and have kept them closed to people who lacked the requisite connections, wealth, or other in-group identifiers. Universities, the gateways to the professions, have long histories of excluding people of colour from their ranks (Burrow 2008, 107).

The historic exclusion of anyone other than White men shapes to this day what mainstream White culture imagines when imagining someone fulfilling such roles. Safiya Noble shows that it took years of criticism before search engines adjusted their algorithms so that searching for "engineer" or "doctor" stopped exclusively returning images of White men (Noble 2018). But the underlying bias, on which the algorithms fed, remains. To imagine a machine in a white-collar job is therefore to imagine a White machine.

Third, hierarchies of intelligence and of professional status are of course also hierarchies of power. Consequently, power relations are implicit in the previous two categories. However, it is worth also considering power separately, because power struggles between AI and humans are such a common narrative trope. Alongside the narrative that robots will make humans redundant, an equally well-known narrative is that they will rise up and conquer us altogether (Cave and Dihal 2019). These are both narratives about machines becoming superior to humans: stories in which they become better at every task, leaving humans with nothing to do, from E.M. Forster's 1909 short story 'The Machine Stops' to the Oscar-winning film *WALL-E*, or in which they outwit and subjugate those who built them, as in the *Terminator* film franchise or the film *Ex Machina* (Forster 1909; Stanton 2008; Cameron 1984; Garland 2015). When White people imagine being overtaken by superior beings, those beings do not resemble those races they have framed as inferior. It is unimaginable to a White audience that they will be surpassed by machines that are Black. Rather, it is by superlatives of themselves: hyper-masculine White men like Arnold Schwarzenegger as the Terminator, or hyper-feminine White women like Alicia Vikander as Ava in *Ex Machina*.

This is why even narratives of an AI uprising that are clearly modelled on stories of slave rebellions depict the rebelling AIs as White—for example, in *Blade Runner* (Dihal 2020). The implication of this racialisation is that these machines might genuinely be superior, or are at least worthy adversaries. The use of White bodybuilders such as Arnold Schwarzenegger to play the evil robots suggests this. As Dyer points out, Schwarzenegger's physique suggests "the body made possible by [...] natural

mental superiority. The point after all is that it is built, a product of the application of thought and planning, an achievement" (Dyer 1997, 164). Consequently, for a White technologist or author, to imagine a superior anthropomorphic machine is to imagine a White machine.

In summary, popular conceptions of AI suggest these machines have general intelligence, are capable of professional jobs, and/or are poised to surpass and supplant humanity. In the White imagination, such qualities are strongly associated with Whiteness. It is no surprise, therefore, that in mainstream Western media, such machines are portrayed as White.

3.3 White Utopia

While we believe the attribution to AI of these qualities, so strongly associated with Whiteness, goes a long way to making sense of the racialisation of anthropomorphic intelligent machines, we also want to propose one further hypothesis: that the Whiteness of the machines allows the White utopian imagination to fully exclude people of colour.

One of the most pertinent hopes for artificial intelligence is that it will lead to a life of ease (Cave and Dihal 2019). As a tool that can take over "dirty, dull, or dangerous" jobs, it relieves its owners from work they do not want to do, enabling them to pursue leisure. As critical race theorists have repeatedly pointed out, the leisure currently available to the wealthier classes is disproportionately facilitated by the labour of working-class women of colour (hooks 1992 1997; Rhee 2018). bell hooks shows that the people performing this labour are actively kept invisible, even when the White master and the coloured servant are physically present in the same space. She cites the memoirs of a White heiress who grew up with Black servants in her house: "Blacks, I realized, were simply invisible to most white people, except as a pair of hands offering a drink on a silver tray" (hooks 1992 1997, 168).

As this forced pretence of invisibility shows, interactions with non-White servants are undesirable to the White master: such interactions are almost literally considered a "dirty job". Depictions of people of colour as being dirty and unwashed, eating dirty food, living in the dirt, even of being the colour of excrement have contributed to the development of both the fear of pollution in interactions with people of colour, and the association of Whiteness with cleanliness and purity (Dyer 1997, 75–76). This association has been exacerbated by a long history of propaganda preceding conquest and genocide that portrays the racial other as evoking disgust: as vectors of disease, such as lice or rats, or as a literal plague (Glover 1999, chap. 35; Rector 2014, chap. 3).

The utopia of the White racial frame would therefore rather remove people of colour altogether, even in the form of servants. From the inception of the academic study of science fiction onwards, many critics have pointed out that utopias throughout literary history have been construed on exclusionary, colonialist, and eugenicist premises (Suvin 1979 2016, 179; Jameson 2005, 205; Ginway 2016, 132). In *Astrofuturism*, De Witt Douglas Kilgore shows that mid-twentieth-century American visions of space age utopias are "idealisations ... based on a series of exclusions" (Kilgore 2010, 10): rather than depicting a post-racial or colourblind future, the authors of these utopias simply omit people of colour.

AI offers the possibility of making such racialised utopias real. By virtue of its generality, it is imagined as able to replace all and any unwanted labour—social and cognitive as well as physical (Cave and Dihal 2019), so obviating the need for people of colour in any role. Consequently, as Jennifer Rhee points out, advertisements for real AI such as household robots "are striking in their whiteness": they are aimed at showing white middle-class families an ideal leisurely lifestyle. In doing so, she argues, "the images reserve the luxury of liberation from domestic labor for white women, while erasing the women of color who perform this labor, both within their own homes and in the homes of others" (Rhee 2018, 94).

In some cases, the unsulliedness of this utopia can extend further to exclude all women. Just as people of colour can be associated with offensive physicality, so can women in general, particularly with respect to their reproductive organs. The necessity of sexual intercourse, pregnancy, and childbearing for the continuation of a race that prides itself on rationality and the ability to transcend its physicality is an offensive hurdle that has been imagined as transcendable by science for centuries. As Dyer points out, in the ideology of Whiteness, the elevation of mental over physical prowess has simultaneously been the White race's most valuable achievement and a threat to its own continuation (Dyer 1997, 27). It has led to the paradox known as the "White Crisis", in which the White race is seen as under threat of being overwhelmed by "inferior" races that are breeding more prolifically. Transhumanism has been envisioned as a solution to this White Crisis (Ali 2017). Seen as a form of offspring, artificial intelligence offers a way for the White man to perpetuate his existence in a rationally optimal manner, without the involvement of those he deems inferior.

4 Conclusion and Implications

Images of AI are not generic representations of human-like machines, but avatars of a particular rank within the hierarchy of the human. These representations of intelligent machines—and our future with them—are refracted through the White racial frame; their Whiteness a proxy for how we perceive their status and potential. This can cause what is sometimes called representational harms (Blodgett et al. 2020). We suggest three.

First, this racialisation can amplify the very prejudices it reflects. We have argued that intelligent machines are portrayed as White because that is how the mainstream perceives intelligence and related desirable characteristics. But equally, the consistent portrayal of intelligent machines as White itself transmits this association, so sustaining it. As we have argued elsewhere (Whittlestone et al. 2019), bias in representations of AI contributes to a vicious cycle of social injustice: the biased representations can influence both aspiring technologists and those in charge of hiring new staff, shaping whom they consider fit for the field (Cave 2020). This could contribute to sustaining a racially homogenous workforce, which will continue to produce products, whether real intelligent machines or their representations, that are biased to benefit that group and disadvantage others.

Second, the racialisation of these machines places them within an existing hierarchy of the human in a way that could exacerbate real injustice. Portrayals of AI as White situate these machines in a power hierarchy above currently marginalised groups, such as people of colour. These oppressed groups are therefore relegated to an even lower

position in the hierarchy: below that of the machine. As machines become ever more important in making automated decisions—frequently about marginalised groups (Eubanks 2017)—this could be highly consequential. Automation bias—the tendency of people to favour suggestions from automated decision-making systems over those from humans—has already been evidenced (Goddard et al. 2012). We might speculate that it will be exacerbated in cases where such systems are racialised White and the humans in question are not.

Third, these portrayals could distort our perceptions of the risks and benefits of these machines. For example, they could frame the debate about AI's impact disproportionately around the opportunities and risks posed to White middle-class men (Cave 2020). It is already a common narrative that the current wave of automation differs from those of the past in that "impacts from automation have thus far impacted mostly blue-collar employment; the coming wave of innovation threatens to upend white-collar work as well" (Pew Research Center 2014). Public interest and policy therefore often focus on white collar professionals, instead of on marginalized groups, which in reality are likely to be worse affected by the impact of AI (Eubanks 2017; Noble 2018).

In this paper, we have offered three interpretations of the whiteness and Whiteness of representations of AI. All three, and the implications that we posit, need further investigation. This process is part of what can be described as *decolonising AI*: a process of breaking down the systems of oppression that arose with colonialism and have led to present injustices that AI threatens to perpetuate and exacerbate. Weheliye describes how he "works towards the abolition of Man, and advocates the radical reconstruction and decolonization of what it means to be human" (Weheliye 2014, 4). It is in the field of AI that technology is most clearly entwined with notions of "what it means to be human", both in reality and in cultural fantasies. We hope to have taken a step towards this reconstruction, by drawing attention to the Whiteness of these machines and "making it strange".

Acknowledgements The authors would like to thank Ezinne Nwankwo, Dr. Lauren Wilcox, Eva Pasini, and the two anonymous peer reviewers for comments on earlier drafts.

Funding Information 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.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Disclaimer The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Templeton World Charity Foundation, Inc.

References

Adas, M. (1989). *Machines as the measure of men: science, technology, and ideologies of Western dominance*. Ithaca: Cornell University Press. https://doi.org/10.7591/9780801455261.

Addison, A., Bartneck, C., and Yogeeswaran, K. (2019). 'Robots can be more than black and white: examining racial bias towards robots'. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society - AIES '19*, 493–98. Honolulu, HI: ACM Press. https://doi.org/10.1145/3306618.3314272.

Ali, S. M. (2017) Transhumanism and/as Whiteness. In *Proceedings of the IS4SI 2017 Summit Digitalisation for a Sustainable Society*. Gothenburg: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/IS4SI-2017-03985.

Barshay, R. (1974). Ethnic stereotypes in "Flash Gordon"'. *Journal of Popular Film, 3*(1), 15–30.

Bartneck, C., Yogeeswaran, K., Ser, Q. M., Woodward, G., Sparrow, R., Wang, S., and Eyssel, F. (2018). 'Robots and racism'. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 196–204. New York: ACM. https://doi.org/10.1145/3171221.3171260.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the New Jim Code. Medford, MA: Polity.

Blodgett, S. L., Barocas, S., Daumé III, H. and Wallach, H. (2020). 'Language (technology) is power: a critical survey of "Bias" in NLP'. ArXiv:2005.14050 [Cs], May 2020. http://arxiv.org/abs/2005.14050.

Buolamwini, J., and Gebru, T. (2018). 'Gender shades: intersectional accuracy disparities in commercial gender classification'. In *Proceedings of Machine Learning Research*. Vol. 81.

Burrow, G. N. (2008). *A history of Yale's School of Medicine: passing torches to others*. New Haven: Yale University Press.

Cameron, J. (1984). *The terminator*. Orion Pictures.

Carson, J. (2006). *The measure of merit: Talents, intelligence, and inequality in the French and American republics, 1750–1940*. Princeton, NJ: Princeton University Press.

Cave, S. (2020). 'The problem with intelligence: its value-laden history and the future of AI'. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*. New York: ACM Press.

Cave, S., & Dihal, K. (2019). 'Hopes and fears for intelligent machines in fiction and reality'. *Nature Machine Intelligence, 1*(2), 74–78. https://doi.org/10.1038/s42256-019-0020-9.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). 'The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals'. *Journal of Personality and Social Psychology, 83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314.

Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*. New York, NY: Basic Books.

Daniels, J. (2013). 'Race and racism in internet studies: a review and critique.' *New Media & Society, 15*(5), 695–719. https://doi.org/10.1177/1461444812462849.

Daniels, J. (2015). '"My brain database doesn't see skin color": color-blind racism in the technology industry and in theorizing the web.' *American Behavioral Scientist, 59*(11), 1377–1393. https://doi.org/10.1177/0002764215578728.

de la Peña, C. (2010). 'The history of technology, the resistance of archives, and the whiteness of race'. *Technology and Culture, 51*(4), 919–937.

Dihal, K. (2020). 'Enslaved minds: artificial intelligence, slavery, and revolt'. In *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*, edited by Stephen Cave, Kanta Dihal, and Sarah Dillon, 189-212. Oxford: Oxford University Press.

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

Eubanks, V. (2017). *Automating inequality: how high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Feagin, J. R. (2006). *Systemic racism: a theory of oppression*. New York: Routledge.

Feagin, J. R. (2013). *The white racial frame: centuries of racial framing and counter-framing* (2nd ed.). New York: Routledge.

Ford, M. (2015). *The rise of the robots: technology and the threat of mass unemployment*. London: Oneworld Publications.

Forster, E. M. 1909. 'The machine stops'. *The Oxford and Cambridge Review*, November 1909. http://archive.ncsa.illinois.edu/prajlich/forster.html.

Frankenberg, R. (Ed.). (1997a). *Displacing whiteness: essays in social and cultural criticism*. Durham, NC: Duke University Press. https://doi.org/10.1215/9780822382270.

Frankenberg, R. (1997b). 'Introduction: local whitenesses, localizing whiteness'. In *Displacing whiteness*, edited by Ruth Frankenberg, 1–33. Durham, NC: Duke University Press. https://doi.org/10.1215/9780822382270-001.

Freedman, A. (2019). 'If you prick Watto, does he not bleed?' *Jewish Currents* (blog). 14 June 2019. https://jewishcurrents.org/if-you-prick-watto-does-he-not-bleed/.

Ganley, T. (2003). 'What's all this talk about whiteness?' *Dialogue, 1*(2), 12–30.

Garland, A. (2015). *Ex Machina*. Universal Pictures.

Garner, S. (2007). *Whiteness: an introduction*. Abingdon: Routledge.

Ginway, M. E. (2016). 'Monteiro Lobato's O Presidente Negro (The Black President): Eugenics and the corporate state in Brazil'. In *Black and Brown planets: The politics of race in science fiction*, edited by Isiah Lavender III, 131–45. Jackson, MI: University Press of Mississippi.

Glover, J. (1999). *Humanity: a moral history of the twentieth century*. London: Jonathan Cape.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). 'Automation bias: a systematic review of frequency, effect mediators, and mitigators.' *Journal of the American Medical Informatics Association: JAMIA, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089.

Gottfredson, L. S. (1997). 'Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography.' *Intelligence 24*(1), 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Haraway, D. J. (1991). *Simians, cyborgs, and women: the reinvention of nature*. London: Free Association Books.

Harding, S. G. (1986). *The science question in feminism*. Ithaca: Cornell University Press.

hooks, b. [1992] (1997). 'Representing whiteness in the black imagination'. In *Displacing whiteness*, edited by Ruth Frankenberg, 165–79. Durham, NC: Duke University Press. https://doi.org/10.1215/9780822382270-006.

Jameson, F. (2005). *Archaeologies of the future: the desire called utopia and other science fictions*. London: Verso.

Keller, E. F. (1985). *Reflections on gender and science*. New Haven: Yale University Press.

Kipling, R. (1899). 'The white man's burden'. *The Times*, 4 February 1899. http://www.kiplingsociety.co.uk/rg_burden1.htm.

Lavender III, I. (2011). *Race in American science fiction*. Bloomington: Indiana University Press.

Liao, Y., and He, J. (2020). 'The racial mirroring effects on human-agent in psychotherapeutic conversation'. *Proceedings of the 25th international conference on intelligent user interfaces (IUI'20).*

Marino, M. (2014). 'The racial formation of chatbots.' *CLCWeb: Comparative Literature and Culture* 16 (5). https://doi.org/10.7771/1481-4374.2560.

Murphy, R. (2000). *Introduction to AI robotics*. Cambridge, MA: MIT Press.

Nakamura, L. (2006). 'Cultural difference, theory and cyberculture studies'. In *Critical Cyberculture Studies*, edited by David Silver and Adrienne Massanari, 29–36. NYU Press.

Noble, S. U. (2018). *Algorithms of oppression: how search engines reinforce racism*. New York: New York University Press.

Okeke, N. A., Howard, L. C., Kurtz-Costes, B., & Rowley, S. J. (2009). 'Academic race stereotypes, academic self-concept, and racial centrality in African American youth'. *Journal of Black Psychology, 35*(3), 366–387. https://doi.org/10.1177/0095798409333615.

Pew Research Center (2014). 'AI, Robotics, and the Future of Jobs'. http://www.pewinternet.org/2014/08/06/future-of-jobs/.

Rector, J. M. (2014). *The objectification spectrum: understanding and transcending our diminishment and dehumanization of others*. Oxford: Oxford University Press.

Rhee, J. (2018). *The robotic imaginary: the human and the price of dehumanized labor*. Minneapolis: University of Minnesota Press.

Riek, L., and Howard, D. (2014). 'A code of ethics for the human-robot interaction profession'. In *Proceedings of We Robot*. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2757805.

Saini, A. (2019). *Superior: The Return of Race Science*. London: 4th Estate.

Sinclair, B. (2004). 'Integrating the histories of race and technology.' In *Technology and the African-American Experience: Needs and Opportunities for Study*, edited by Bruce Sinclair, 1-17. Cambridge, MA: MIT Press.

Stanton, A. (2008). *WALL·E*. Disney. http://www.imdb.com/title/tt0910970/.

Strait, M., Ramos, A. S., Contreras, V., and Garcia, N. (2018). 'Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white'. In *The 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 452–57. https://doi.org/10.1109/ROMAN.2018.8525610.

Suvin, D. [1979] (2016). *Metamorphoses of science fiction: on the poetics and history of a literary genre*. Edited by Gerry Canavan. Bern: Peter Lang.

Sweeney, M. (2016). 'The Ms. Dewey "experience:" technoculture, gender, and race'. *Digital Sociologies*. https://doi.org/10.2307/j.ctt1t89cfr.31.

UNDP (2017). 'UNDP in Asia and the Pacific appoints world's first non-human innovation champion'. *UNDP in Asia and the Pacific*. 22 November 2017. https://www.asia-pacific.undp.org/content/rbap/en/home/presscenter/pressreleases/2017/11/22/rbfsingapore.html.

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). 'The implicit prejudiced attitudes of teachers: relations to teacher expectations and the ethnic achievement gap.' *American Educational Research Journal, 47*(2), 497–527. https://doi.org/10.3102/0002831209353594.

Villa-Nicholas, M., and Sweeney, M. E. (2019). 'Designing the "good citizen" through Latina identity in USCIS's virtual assistant "Emma"'. *Feminist Media Studies*, July, 1–17. https://doi.org/10.1080/14680777.2019.1644657.

Weheliye, A. G. (2014). *Habeas Viscus: racializing assemblages, biopolitics, and black feminist theories of the human*. Durham, NC: Duke University Press.

Weller, C. (2017). 'Meet Sophia, the robot citizen that said it would "destroy humans"'. *Business Insider*. 27 October 2017. https://www.businessinsider.com/meet-the-first-robot-citizen-sophia-animatronic-humanoid-2017-10?r=UK.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., and Cave, S. (2019). 'Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research'. Nuffield Foundation.

Williams, P. J. (1999). 'Racial Ventriloquism'. *The Nation*, 17 June 1999. https://www.thenation.com/article/archive/racial-ventriloquism/.

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*(3), 257–337.