Racialization and Bias toward Humanoids

Kumar Yogeeswaran, Christoph Bartneck

December 14, 2024

Abstract

Humans categorize themselves and others on the basis of many attributes forging a range of social groups. Such group identities can influence our perceptions, attitudes, beliefs, and behaviors toward others and ourselves. While decades of psychological research has examined how dividing the world into 'us' and 'them' impacts our attitudes, beliefs, and behaviors toward others, a new and emerging area of research considers how humans can ascribe social group memberships to humanoid robots. Specifically, our social perceptions and evaluations of humanoids can be shaped by subtle characteristics about the robot's appearance or other features, particularly if these characteristics are interpreted through the lens of important human group identities. The current chapter reviews research on the psychology of intergroup relations to consider its manifestations and expressions in the context of human-robot interaction. We first consider how robots despite being non-living can be ascribed certain identities (e.g., race, gender, national origin). We then consider how this can in turn impact attitudes, beliefs, and behaviors toward such technology. Given the nascency of this field of study, we highlight existing gaps in our knowledge and highlight important directions for future research. The chapter concludes by considering the societal, market, and legal implications of bias in the context of human-robot interaction.

Address And Affiliation

Kumar Yogeeswaran

Private Bag 4800 School of Psychology, Speech and Hearing University of Canterbury Christchurch 8140 New Zealand Email: kumar.yogeeswaran@canterbury.ac.nz

Christoph Bartneck

Private Bag 4800 Department of Computer Science and Software Engineering University of Canterbury Christchurch 8140 New Zealand Email: christoph.bartneck@canterbury.ac.nz

Bio

Kumar Yogeeswaran

Kumar Yogeeswaran is an Associate Professor of Social Psychology at the University of Canterbury, New Zealand. He completed his PhD in Social Psychology with a concentration in Statistics at the University of Massachusetts - Amherst, USA in 2012. His research broadly examines how people's membership in social groups impacts on their thoughts, feelings, and behaviors toward others, as well as their self-conceptions. As a primary focus, Kumar's research examines intergroup relations and conflict in pluralistic nations comprised of ethnic, racial, religious, and ideological diversity. As a secondary focus, Kumar conducts research that applies psychological science to the fields of human-robot interaction, social media communications, and politics. He utilizes data gathered through varied

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

research instruments including reaction-time, brain-imaging, social media, self-report, and behavioral measures. His research is conducted in both the lab and field using experimental, correlational, and longitudinal designs through convenience, community, and nationally representative samples from several countries. To date, he has published over 80 peer-reviewed scientific articles, including work in leading peer-reviewed journals such as Psychological Science, Journal of Personality and Social Psychology, and Psychological Review.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

Christoph Bartneck

Dr. Christoph Bartneck is a Professor in the Department of Computer Science and Software Engineering at the University of Canterbury. He has a background in Industrial Design and Human-Computer Interaction, and his projects and studies have been published in leading journals, newspapers, and conferences. His interests lie in the fields of Human-Computer Interaction, Science and Technology Studies, and Visual Design. More specifically, he focuses on the effect of anthropomorphism on human-robot interaction. As a secondary research interest he works on projects in the area of sport technology and the critical review on scientific processes and policies. In the field of Design Christoph investigates the history of product design, tessellations and photography. He has worked for several international organizations including the Technology Centre of Hannover (Germany), LEGO (Denmark), Eagle River Interactive (USA), Philips Research (Netherlands), ATR (Japan), and The Eindhoven University of Technology (Netherlands). Christoph is an associate editor of the International Journal of Social Robotics. Christoph is a member of the ACM SIGCHI, The New Zealand Association Of Scientists and Academic Freedom Aotearoa. The press regularly reports on his work, including the New Scientist, Scientific American, Popular Science, Wired, New York Times, The Times, BBC, Huffington Post, Washington Post, The Guardian, and The Economist.

1 Introduction

Human beings belong to a range of social groups and can be differentiated from one another on the basis of many identities including one's age, religion, race, ethnicity, nationality, gender, political ideology, sexual orientation, or social class. These differing identities can shape our perceptions, attitudes, beliefs, and behaviors about another person on a daily basis. However, our social behaviors and perceptions can also be influenced by various other factors about a person's appearance even if these are not tied to a distinct group identity. For example, people's attitudes, beliefs, and social perceptions can be influenced by mental associations we have about others on the basis of their hair color, skin color, eye color, whether they have tattoos or body piercings, their weight, physical attractiveness, etc. As a concrete case in point, studies reveal that people perceive physically attractive individuals as possessing a range of positive attributes including the belief that they are socially competent, intelligent, successful, and well-adjusted (for a review, see Eagly et al.[1]). Similarly, studies also reveal that physically attractive individuals tend to be allocated higher pay for the same quality work as a less attractive counterpart, and in legal contexts, attractive criminals tend to receive lighter sentences and smaller fines than less attractive criminals.[2] Such work illustrates that our attitudes, beliefs, and behaviors can can be clouded by various aspects of another person, and these may even occur without people's conscious awareness. In the current chapter, we explore how our social perceptions and evaluations of humanoid robots (i.e., robots that appear human-like) can be shaped by subtle characteristics about the robot's appearance, especially if these characteristics can be interpreted through the lens of important human group identities.

[1] Alice Eagly et al. "What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype". In: *Journal of Personality and Social Psychology* 110 (1991), pp. 109–128. DOI: 10.1037/0033-2909.110.1.109. URL: https://doi.org/10.1037/0033-2909.110.1.109.

[2] Rosemarie Anderson and Steve A. Nida. "Effect of physical attractiveness on opposite- and same-sex evaluations". In: *Journal of Personality* 46.3 (1978), pp. 401–413. DOI: 10.1111/j.1467-6494.1978.tb01008.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.1978.tb01008.x; John E. Stewart II. "Defendant's Attractiveness as a Factor in the Outcome of Criminal Trials: An Observational Study1". In: *Journal of Applied Social Psychology* 10.4 (1980), pp. 348–361. DOI: 10.1111/j.1559-1816.1980.tb00715.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1980.tb00715.x; Irene Hanson Frieze, Josephine E. Olson, and June Russell. "Attractiveness and Income for Men and Women in Management1". In: *Journal of Applied Social Psychology* 21.13 (1991), pp. 1039–1057. DOI: 10.1111/j.1559-1816.1991.tb00458.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.1991.tb00458.x; Stephen G. West and T. Jan Brown. "Physical attractiveness, the severity of the emergency and helping: A field experiment and interpersonal simulation". In: *Journal of Experimental Social Psychology* 11.6 (1975), pp. 531–538. ISSN: 0022-1031. DOI: 10.1016/0022-1031(75)90004-9. URL: https://www.sciencedirect.com/science/article/pii/0022103175900049.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

2 Intergroup Differences in the Context of Race

Human beings evolved as social creatures that depend on each other for survival. We rely on others for reproduction, gathering food, protection, taking care of us when we are sick, among other foundational needs. The human desire to be part of social groups can be motivated by a desire to manage uncertainty,[3] fulfil our need to belong,[4] and provide us with a sense of positive distinctiveness and social identity that allows us to feel valued and unique.[5] Psychologists argue that we strive for optimal distinctiveness by belonging to social groups that fulfil the above needs, but we simultaneously wish to be part of groups that help us feel unique, special, and distinctive.[6] This is why we may not feel content being identified at the most superordinate level as simply fellow human beings while ignoring various other identities. But people also, typically, do not disregard all group affiliations and become content identifying as only a unique individual with no social affiliations. While shared connection with others offers great benefits to the individual including to our health and well-being, and maximizes our chance at survival by improving our ability to collectively problem solve and share resources, it can also lead to an 'us' and 'them' mind set. Across the world and throughout human history, different aspects of what defines 'us' and 'them' have been the basis for intergroup conflict.

Among these many foundations for dividing ourselves is the concept of race that in some circumstances gives way to racism. While anthropologists and biologists suggest there is no biological basis for race,[7] broad conceptualizations of race have been used to divide people into distinct groups and create systems that favour one over another. For example, conceptions of race at various points of US history were used to legitimize the subjugation of African Americans, Native Americans, among other non-White peoples. Similarly, South Africa maintained an apartheid system until the 1990s that prevented the mixing of Black, Indian, and White South Africans and prevented non-White South Africans from pursuing economic, education, and political advancement. While legal and structural barriers have long persisted and facilitated negative outcomes for many groups throughout history, the exploration of such influences are better captured by a focus on systemic biases and barriers that can create and perpetuate inequalities. However, racism and its consequences can also emerge through people's beliefs and attitudes about others based on their group membership. Such biases may not only reflect people's conscious and overt beliefs and attitudes toward others, but even unconscious or automatic forms of the same. The emergence of new anthropomorphic humanoid technologies raises questions about how people will come to see and interact with such technology that has human-like features. Do we perceive humanoids as competent or warm? Do we perceive humanoids as having a race or gender? Do we apply stereotypes and prejudices we have toward humans social groups to technology that resembles humans? The current chapter explores these questions.

3 Blatant to Subtle Forms of Bias

People's attitudes and beliefs about others on the basis of varied group memberships (e.g., their race, religion, gender, nationality, etc.) may be captured through conscious level responses people may have about an individual. For example, people can self-report that they believe (non)religious people are

[3] Michael A. Hogg. "Uncertainty–Identity Theory". In: Advances in Experimental Social Psychology 39 (2007), pp. 69–126. ISSN: 0065-2601. DOI: 10.1016/S0065-2601(06)39002-8. URL: https://www.sciencedirect.com/science/article/pii/S0065260106390028.

[4] Roy Baumeister and Mark Leary. "The need to belong: Desite for interpersonal attachments as a fundamental human motivation". In: *Psychological Bulletin* 117 (1995), pp. 497–529. URL: https://doi.org/10.1037/0033-2909.117.3.497.

[5] Henri Tajfel and John C Turner. "The social identity theory of intergroup behavior". In: *Psychology of Intergroup Relations*. Nelson Hall, 1986, pp. 7–24.

[6] Marilynn B Brewer. "The social self: On being the same and different at the same time". In: *Personality and social psychology bulletin* 17.5 (1991), pp. 475–482. DOI: 10.1177/014616729117500. URL: https://doi.org/10.1177/014616729117500.

[7] Audrey Smedley and Brian D Smedley. "Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race." In: *American psychologist* 60.1 (2005), pp. 16–26. DOI: 10.1037/0003-066X.60.1.16. URL: https://doi.org/10.1037/0003-066X.60.1.16.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

trustworthy when asked in an interview, or when asked via a questionnaire using a Likert-style measure with a 7-point scale. Such conscious level responses have provided valuable insight into how people think, feel, and may behave toward others on the basis of various important dimensions. However, our interactions with the world can involve two different kinds of processing: system 1 thinking that involves automatic, implicit, or unconscious processing,[8] or it can involve system 2 thinking that involves deliberate, explicit, or conscious-level processing. A large body of research now reveals that human biases toward other social groups can emerge at both the implicit or explicit level (for reviews, see Dasgupta; Greenwald and Lai; and Yogeeswaran, Devos, and Nash[9]).

While genocide, intergroup violence, and blatant dehumanization still occur in some parts of the world, the prevalence of such bias and hostility have subsided in recent history.[10] Nevertheless, biases persist to this day. Social norms, especially within many liberal democracies which make it relatively unacceptable to express blatant animosity toward others based on their group membership. However, people may subtly treat others unfairly in the form of everyday biases, or even without full awareness of the actor. Therefore, there has been an increased focus among social scientists to use different tools to better understand bias that emerge in subtler ways. For example, decades of research on aversive racism has examined how even well-intentioned people with egalitarian values may discriminate against outgroup members when their behavior is easy to rationalize (for a review, see Dovidio, Gaertner, and Pearson[11]). For example, in a classic study on the topic,[12] researchers measured the self-reported attitudes of two independent samples of White Americans in 1989 and 1999 using a questionnaire. In parallel, these participants at both time points were asked to evaluate a job candidate who was either African American or White American for an identical role, only they further manipulated the qualifications of the African American or White candidate such that the individual would be presented as either highly qualified for the role, ambiguously qualified (i.e., there were aspects of their qualifications that one could use to make a case for or against them), or they were weakly qualified for the role. All participants then indicated the extent to which they would recommend hiring the candidate for the specific role. First, data revealed that between 1989 and 1999, there was a decrease in self-reported negative attitudes toward African Americans relative to White Americans suggesting that there was a decrease in self-reported or explicit prejudice toward African Americans over the span of a decade. However, when it came to people's recommendations for hiring, there was no difference in their evaluations in 1989 and 1999. Specifically, data revealed that when the candidate was highly qualified for the job, participants were equally likely to recommend both the African American and White American candidate. Similarly, when the candidate was weakly qualified, participants were equally likely to oppose both the African American and White American candidate for the role. However, it was specifically when the candidate was ambiguously qualified, the White American candidate was more likely to be recommended for the role than the African American candidate. In such a situation, it would be easier to rationalize one's decision in either direction allowing people who endorse egalitarian values to discriminate.[13] Such work reveals the importance of going beyond self-report tools to examine intergroup biases via more indirect tools of measurement.

[8] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[9] Nilanjana Dasgupta. "Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations". In: *Social justice research* 17.2 (2004), pp. 143–169. DOI: 10.1023/B:SORE.0000027407.70241.15. URL: https://doi.org/10.1023/B:SORE.0000027407.70241.15; Anthony G Greenwald and Calvin K Lai. "Implicit social cognition". In: *Annual Review of Psychology* 71.1 (2020), pp. 419–445. DOI: 10.1146/annurev-psych-010419-050837. URL: https://doi.org/10.1146/annurev-psych-010419-050837; Kumar Yogeeswaran, Thierry Devos, and Kyle Nash. "Understanding the nature, measurement, and utility of implicit intergroup biases". In: *The Cambridge handbook of the psychology of prejudice*. New York, NY, US: Cambridge University Press, 2017, pp. 241–266. ISBN: 978-1-107-09833-6 (Hardcover). DOI: 10.1017/9781316161579.011. URL: https://doi.org/10.1017/9781316161579.011.

[10] Steven Pinker. *Enlightenment now: The case for reason, science, humanism, and progress*. Penguin UK, 2018.

[11] John F. Dovidio, Samuel L. Gaertner, and Adam R. Pearson. "Aversive racism and contemporary bias". In: *The Cambridge handbook of the psychology of prejudice*. New York, NY, US: Cambridge University Press, 2017, pp. 267–294. ISBN: 978-1-107-09833-6 (Hardcover). DOI: 10.1017/9781316161579.012. URL: https://doi.org/10.1017/9781316161579.012.

[12] John F Dovidio and Samuel L Gaertner. "Aversive racism and selection decisions: 1989 and 1999". In: *Psychological science* 11.4 (2000), pp. 315–319. DOI: 10.1111/1467-9280.0026. URL: https://doi.org/10.1111/1467-9280.0026.

[13] Ibid.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

One area of research that has considerably expanded our understanding of subtle biases comes from research in the field of implicit social cognition. For several decades, using insights initially developed in cognitive psychology, social psychologists have been using a range of tools to understand implicit biases. For example, a large number of studies rely on reaction-time measures focusing on either latency (i.e., how long people take to associate certain attributes or positive vs. negative valence with one group over another), or error rates (i.e., how many errors we make when rapidly trying to categorize certain attributes or positive vs. negative concepts with one group over another) to capture implicit bias. These tools allow us to better understand the strength of association between two concepts by better understanding how easily these associations come to mind (for reviews, see Dasgupta, Greenwald and Lai, and Gawronski and Hahn[14]). These reaction-time tools come in different forms (e.g., the Implicit Association Task, evaluative priming, Go/No-Go Association Task) with varied costs and benefits (for a review on these measures, see[15]). However, recent research has called into question the extent to which these implicit measures capture unconscious attitudes or beliefs about other groups as people are surprisingly accurate in predicting their implicit bias scores[16] and others raise concerns about the construct validity of these measures as indexes of individual differences in bias.[17] Nevertheless, responses on self-report and reaction-time measures are often weakly correlated or entirely unrelated suggesting both capture distinct aspects of people's evaluations and beliefs about social groups.

Among the different reaction-time measures developed in the literature to assess these relatively automatic or implicit biases is a shooter bias task. Correll et al.[18] initially developed the shooter bias paradigm following the tragic shooting of Amadou Diallo, an unarmed African American man, in 1999. Diallo was shot 19 times by 4 New York police officers who mistakenly assumed that Diallo was reaching out to grab a weapon from his pocket which turned out to be his wallet. Correll et al.[19] used a series of images involving African American and White American men holding guns and innocuous objects (e.g., a wallet, soda can, cellphone) in different poses and across various backgrounds. Participants were asked to rapidly respond by indicating a 'shoot' or 'don't shoot' response to each of these trials. Based on either the number of errors participants made (i.e., how often they mistakenly shoot unarmed Black vs. White targets, or don't shoot armed Black vs. White targets), or the time they took to respond to each of these accurately, a shooter bias score was calculated. Across dozens of studies, there is evidence that people engage in shooter bias against African American over White American men. A recent meta-analysis combining many studies in the literature together provides further evidence of shooter bias.[20] This bias is found even among African American participants,[21] and seems to be driven by awareness of the cultural stereotypes about the group.[22] Moreover, this bias appears particularly

[14]Dasgupta, "Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations"; Greenwald and Lai, "Implicit social cognition"; Bertram Gawronski and Adam Hahn. "Implicit measures: Procedures, use, and interpretation". In: *Measurement in social psychology*. Routledge, 2018, pp. 29–55.

[15]Russell H Fazio and Michael A Olson. "Implicit measures in social cognition research: Their meaning and use". In: *Annual review of psychology* 54.1 (2003), pp. 297–327. DOI: 10.1146/annurev.psych.54.101601.145225. URL: https://doi.org/10.1146/annurev.psych.54.101601.145225; Yogeeswaran, Devos, and Nash, "Understanding the nature, measurement, and utility of implicit intergroup biases".

[16]Adam Hahn et al. "Awareness of implicit attitudes". In: *Journal of Experimental Psychology: General* 143.3 (2014), pp. 1369–1392. DOI: 10.1037/a0035028. URL: https://doi.org/10.1037/a0035028.

[17]Ulrich Schimmack. "The Implicit Association Test: A method in search of a construct". In: *Perspectives on Psychological Science* 16.2 (2021), pp. 396–414. DOI: 10.1177/174569161986379. URL: https://doi.org/10.1177/174569161986379.

[18]Joshua Correll et al. "The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals". In: *Journal of Personality and Social Psychology* 83.6 (2002), pp. 1314–1329. ISSN: 0022-3514. DOI: 10.1037//0022-3514.83.6.1314. URL: http://doi.org/10.1037/0022-3514.83.6.1314.

[19]Ibid.

[20]Yara Mekawi and Konrad Bresin. "Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis". In: *Journal of Experimental Social Psychology* 61 (2015), pp. 120–130. DOI: 10.1016/j.jesp.2015.08.002. URL: https://doi.org/10.1016/j.jesp.2015.08.002.

[21]Correll et al., "The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals".

[22]Joshua Correll et al. "The influence of stereotypes on decisions to shoot". In: *European Journal of Social Psychology* 37.6 (2007), pp. 1102–1117. DOI: 10.1002/ejsp.450. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.450; Joshua Correll, Geoffrey R Urland, and Tiffany A Ito. "Event-related potentials and the decision to shoot: The role of threat perception and cognitive control". In: *Journal of Experimental Social Psychology* 42.1 (2006), pp. 120–128.

in threatening contexts such as a dangerous neighbourhood or when targets wear clothing perceived as threatening versus when targets were in neighbourhoods or with clothing perceived to be safe.[23] However, it is worth noting that this bias seems especially targeted toward African American men over African American women, White men, and White women suggesting that the bias reflects a behavioral threat-related response particularly directed to outgroup men who are stereotypically associated with aggression.[24] This is further reinforced by the finding that Asian targets are less likely to experience shooter bias than even White targets.[25] Interestingly, this bias is less prevalent among highly trained police officers (i.e., SWAT team officers) relative to the general population,[26] and training that weakens the association between race and danger can attenuate these biases.[27]

4 Utility of Implicit Measures

While the literature on implicit bias continues to evolve and develop both theoretically and empirically, a fundamental question that researchers have been interested in since these tools were developed is whether people's responses on these implicit measures translate into any meaningful outcomes. For more than two decades, psychologists have debated the merit and utility of these implicit biases. Such research has sought to examine whether implicit attitudes and beliefs predict real-life behavior in a range of contexts including job-hiring, voting, medical decision-making, etc. A 2009 meta-analysis[28] using 184 independent samples first examined the relationship between implicit and explicit attitudes and beliefs with behavioral, judgement, and physiological measures to suggest that both implicit and self-report tools were useful in predicting the above outcomes (see also Kurdi et al.[29]). On socially sensitive topics such as in prejudice and racism related domains, implicit tools were shown to be even more useful than self-report. However, a 2013 meta-analysis[30] using 27 of the 32 publications reported in the[31] meta-analysis, along with an additional 18 reports published between 2007 and 2011 estimated a weaker relationship between implicit measures and discrimination. Specifically, the relationship between implicit and explicit attitudes and beliefs specifically with discrimination across a range of domains revealed that implicit measures were poorer predictors of discrimination across

DOI: 10.1016/j.jesp.2005.02.006. URL: https://doi.org/10.1016/j.jesp.2005.02.006.

[23] Kimberly Barsamian Kahn and Paul G Davies. "What influences shooter bias? The effects of suspect race, neighborhood, and clothing on decisions to shoot". In: *Journal of Social Issues* 73.4 (2017), pp. 723–743. DOI: 10.1111/josi.12245. URL: https://doi.org/10.1111/josi.12245.

[24] Ashby Plant, Joanna Goplen, and Jonathan Kuntsman. "Selective responses to threat: The roles of race and gender in decisions to shoot". In: *Personality and Social Psychology Bulletin* 37.9 (2011), pp. 1274–1281. DOI: 10.1177/0146167211408617. URL: https://doi.org/10.1177/0146167211408617.

[25] Melody S Sadler et al. "The world is not black and white: Racial bias in the decision to shoot in a multiethnic context". In: *Journal of Social Issues* 68.2 (2012), pp. 286–313. DOI: 10.1111/j.1540-4560.2012.01749.x. URL: https://doi.org/10.1111/j.1540-4560.2012.01749.x.

[26] Joshua Correll et al. "Across the thin blue line: police officers and racial bias in the decision to shoot". In: *Journal of personality and social psychology* 92.6 (2007), pp. 1006–1023. DOI: 10.1037/0022-3514.92.6.1006. URL: https://doi.org/10.1037/0022-3514.92.6.1006.

[27] Correll et al., "Across the thin blue line: police officers and racial bias in the decision to shoot"; E Ashby Plant and B Michelle Peruche. "The consequences of race for police officers' responses to criminal suspects". In: *Psychological Science* 16.3 (2005), pp. 180–183. DOI: 10.1111/j.0956-7976.2005.008. URL: https://doi.org/10.1111/j.0956-7976.2005.008; Jessica J Sim, Joshua Correll, and Melody S Sadler. "Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot". In: *Personality and social psychology bulletin* 39.3 (2013), pp. 291–304. DOI: 10.1177/0146167212473157. URL: https://doi.org/10.1177/0146167212473157.

[28] Anthony G Greenwald et al. "Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity". In: *Journal of personality and social psychology* 97.1 (2009), pp. 17–41. DOI: 10.1037/a0015575. URL: https://doi.org/10.1037/a0015575.

[29] Benedek Kurdi et al. "Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis". In: *American psychologist* 74.5 (2019), pp. 569–586. DOI: 10.1037/amp0000364. URL: https://doi.org/10.1037/amp0000364.

[30] Frederick L Oswald et al. "Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies". In: *Journal of personality and social psychology* 105.2 (2013), pp. 171–192. DOI: 10.1037/a0032734. URL: https://doi.org/10.1037/a0032734.

[31] Greenwald et al., "Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity".

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

domains and performed worse than self-report measures (also see Forscher et al.[32]). The differences in the findings across these meta-analyses may be partly due to differences in the criteria for inclusion in their respective meta-analyses, but along with a more recent meta-analysis,[33] it would appear that implicit biases have at least a small predictive relationship with discriminatory behavior (but also see Carlsson and Agerström[34]).

In recent years, emerging from research on the 'wisdom of crowds'[35] which suggests that aggregation of information in a group results in better decisions than that made by a single member of the group[36] put forward a theoretical framework for what is referred to as the 'bias of crowds'. According to this framework, rather than focus so much on the extent to which implicit biases within an individual contribute to their own racist and discriminatory behaviors, it is better to consider the aggregate context-level implicit bias as a predictor of real-world systemic racism.[37] Specifically, research from this framework argues that implicit biases are more informative at an aggregate context level (i.e., by looking at differences at the neighbourhood, county, state, or even country level) and explain real-world discrimination and disparities. Proponents of this framework point to several studies showing that variation in levels of implicit bias across communities correlate with real-world inequities and discrimination. For example, cross-national differences in implicit (but not self-reported) gender stereotypes associating male with science and female with liberal arts across 34 countries predicted nation-level sex differences in 8th grade mathematics and science achievement.[38] Similarly, US metropolitan areas with higher levels of implicit bias toward African Americans relative to White Americans were found to have greater disparities in the police shootings of African Americans relative to White Americans.[39] And people with HIV reported higher levels of psychological distress when living in towns with higher level of implicit prejudice toward people with HIV[40] suggesting that even though implicit biases tend to reflect relatively automatic associations, aggregate levels of implicit bias can have downstream implications toward real-world outcomes.

Taken together, the chapter thus far describes how people's perceptions and behaviors can both consciously and relatively unconsciously be influenced by the perceived group membership of others. But are our thoughts, feelings, and behaviors toward inanimate technology such as robots influenced by identities we may apply to them?

[32]Patrick S Forscher et al. "A meta-analysis of procedures to change implicit measures". In: *Journal of personality and social psychology* 117.3 (2019), pp. 522–559. DOI: 10.1037/pspa0000160. URL: https://doi.org/10.1037/pspa0000160.

[33]Kurdi et al., "Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis".

[34]Rickard Carlsson and Jens Agerström. "A closer look at the discrimination outcomes in the IAT literature". In: *Scandinavian journal of psychology* 57.4 (2016), pp. 278–287. DOI: 10.1111/sjop.12288. URL: https://doi.org/10.1111/sjop.12288.

[35]James Surowiecki. "The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business". In: *Economies, Societies and Nations* 296.5 (2004).

[36]B Keith Payne, Heidi A Vuletich, and Kristjen B Lundberg. "The bias of crowds: How implicit bias bridges personal and systemic prejudice". In: *Psychological Inquiry* 28.4 (2017), pp. 233–248. DOI: 10.1080/1047840X.2017.1335568. URL: https://doi.org/10.1080/1047840X.2017.1335568.

[37]Payne, Vuletich, and Lundberg, "The bias of crowds: How implicit bias bridges personal and systemic prejudice"; B Keith Payne and Jason W Hannay. "Implicit bias reflects systemic racism". In: *Trends in cognitive sciences* 25.11 (2021), pp. 927–936. DOI: 10.1016/j.tics.2021.08.001. URL: https://doi.org/10.1016/j.tics.2021.08.001.

[38]Brian A Nosek et al. "National differences in gender–science stereotypes predict national sex differences in science and math achievement". In: *Proceedings of the National Academy of Sciences* 106.26 (2009), pp. 10593–10597. DOI: 10.1073/pnas.0809921106. URL: https://doi.org/10.1073/pnas.0809921106.

[39]Eric Hehman, Jessica K Flake, and Jimmy Calanchini. "Disproportionate use of lethal force in policing is associated with regional racial biases of residents". In: *Social psychological and personality science* 9.4 (2018), pp. 393–401. DOI: 10.1177/1948550617711229. URL: https://doi.org/10.1177/1948550617711229.

[40]Carol T Miller et al. "Macro-level implicit HIV prejudice and the health of community residents with HIV". in: *Health Psychology* 35.8 (2016), pp. 807–815. DOI: 10.1037/hea0000314. URL: https://doi.org/10.1037/hea0000314.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

5 Bias in Human-Robot Interaction (HRI)

As mentioned before, humans can perceive themselves and others through the lens of various group identities and attributes. Self-categorization theory[41] argues that the self can be represented at varying levels of abstraction going from the narrower level of a personal identity (i.e., about who 'I' am, and aspects of myself) to a broader conception of the social self that includes the various social groups we are associated with (e.g., our ethnic, national, racial, religious, political identities), and at the most abstract level, we can represent ourselves as humans where we distinguish ourselves from non-humans (e.g., animals, plants, aliens, or robots). Robots, therefore, represent a distinct outgroup to the human ingroup. An extensive body of research has examined how anthropomorphism of robots by making them more 'human' like impacts on people's thoughts, feelings, and behaviors (for a review, see Zlotowski et al.[42]). However, here we will focus on how we may apply human categories and identities to robots and how such representations of robots can influence our attitudes, judgements, and behaviors toward these. Although people may recognize robots as a distinct social entity from themselves, our evaluations of them may depend on the extent to which they become associated with human social identities. For example, when a robot is perceived as being made in a certain country, we may apply stereotypic beliefs about that country on to our judgements about the technology. For example, a robot that is made in Germany may be perceived as more competent than a robot made in New Zealand based on national stereotypes about competence. Interestingly, the label "Made in Germany" was originally introduced in the UK to protect UK quality products against the cheaper German imports. Unfortunately, the quality of the German products was much better than those from the UK and hence the label transformed from an indicator for inferior foreign products to high quality German products. However, a robot made in Germany may be perceived as less warm than a robot made in New Zealand based on national stereotypes that suggest New Zealand is a friendlier country. The quality of products, or lack thereof, can therefore be ascribed to the robot based on its country of origin.

People may also apply group-based stereotypes about the robot simply based on its physical appearance or its name and voice. The color of the plastic, for example, can be easily changed to different shades of brown to trigger racialization of the robots.[43] This racialization then changes how people interact with the robots. Bartneck et al. and Addison, Yogeeswaran, and Bartneck[44] replicated the classical 'Shooter Bias' study[45] that showed how a racialized robot is treated differently in the context of a police shooting task. Similarly, Strait et al.[46] showed that a robot racialized as Black or Asian is dehumanised more than a White robot. When the racialization is limited to only certain sections of the face and not extended to the whole body, the effect of the racialization is reduced.[47]

There are two more possible interpretations of the term 'race' in the context of HRI that we are

[41] John C Turner et al. *Rediscovering the social group: A self-categorization theory*. basil Blackwell, 1987.

[42] Jakub Zlotowski et al. "Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction". In: *International Journal of Social Robotics* 7.3 (2015), pp. 347–360. DOI: 10.1007/s12369-014-0267-6. URL: https://doi.org/10.1007/s12369-014-0267-6.

[43] Friederike Eyssel and Steve Loughnan. ""It Don't Matter If You're Black or White"?" In: *Proceedings of the 5th International Conference on Social Robotics - Volume 8239*. ICSR 2013. Bristol, UK: Springer-Verlag, 2013, pp. 422–431. ISBN: 9783319026749. DOI: 10.1007/978-3-319-02675-6_42. URL: https://doi.org/10.1007/978-3-319-02675-6_42.

[44] C. Bartneck et al. "Robots and Racism". In: *ACM/IEEE International Conference on Human Robot Interaction*. ACM, 2018, pp. 196–204. DOI: 10.1145/3171221.3171260. URL: https://doi.org/10.1145/3171221.3171260; Arifah Addison, Kumar Yogeeswaran, and C. Bartneck. "Robots Can Be More Than Black And White: Examining Racial Bias Towards Robots". In: *AAAI/ACM Conference On Artifical Intelligence, Ethics, and Society*. ACM, 2019, pp. 493–498. DOI: 10.1145/3306618.3314272. URL: https://doi.org/10.1145/3306618.3314272.

[45] Correll et al., "The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals".

[46] Megan Strait et al. "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)*. 2018, pp. 452–457. DOI: 10.1109/ROMAN.2018.8525610. URL: https://doi.org/10.1109/ROMAN.2018.8525610.

[47] Jaime Banks and Kevin Koban. "A Kind Apart: The Limited Application of Human Race and Sex Stereotypes to a Humanoid Social Robot". In: *International Journal of Social Robotics* (2022). ISSN: 1875-4805. DOI: 10.1007/s12369-022-00900-2. URL: https://doi.org/10.1007/s12369-022-00900-2.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

not going to consider. First, we will not consider the engineers that designed and produced the robots. Their ethnicity, nationality and race are largely independent from the robots' design. Moreover, little information is available on the creators of robots. The second form of 'race' we will not consider is how a robot reacts to the race of a human user. The robots' could exhibit a bias towards certain groups of users. These types of biases are an overall concern for all AI systems. Some facial recognition systems, for example, worked better with Caucasian faces compared to African American faces.[48] Instead, here we consider how people may ascribe race to robots and consider how this may impact their perceptions and behaviors toward such technology.

5.1 Robots can be considered a distinct social category from humans

The term 'race' does not normally apply to artefacts. Although race may be a social construction,[49] it is typically applied to groups of people, while robots are non-living machines. However, people respond to robots similarly to how they respond to humans. This anthropomorphization is a well researched topic in the field of HRI.[50] Social categories, such as age, gender and race are often applied to robots to form opinions.[51] The body shape and other physical properties of robots, for example, can be manipulated to evoke gender stereotypes.[52]

We therefore have to reconsider what we mean with the term 'race' in the context of HRI. One could consider all robots to be a distinct social category, similar to how plants are a distinct social category. However, people may ascribe racial identities to robots even if they are seen as distinct from humans and this racialization of robots may impact people's attitudes, behaviors, and judgements toward these robots. Similarly, people could ascribe social class to a robot based on its perceived occupation or social status, and robots could be classified based on their function, such as floor cleaning robots versus companion robots. If so, then, in the same way that we might have stereotypic beliefs or prejudices toward lawyers or cleaners, we might develop biases towards certain groups of robots that fulfil certain functions.

Which brings us to the last interpretation of the term 'race' in the context of HRI. Robots could be racialised based on their physical properties, such as their facial features or the colour of their plastic surface. Here the racialization could be based on what human racial categories the particular robot elicits. A dark brown robot located in the US may be racialized as Black or African American. The

[48] Jacqueline G. Cavazos et al. "Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias?" In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.1 (2021), pp. 101–111. DOI: 10.1109/TBIOM.2020.3027269. URL: https://doi.org/10.1109/TBIOM.2020.3027269.

[49] Smedley and Smedley, "Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race."

[50] Zlotowski et al., "Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction".

[51] Robert Sparrow. "Do Robots Have Race?: Race, Social Construction, and HRI". in: *IEEE Robotics Automation Magazine* 27.3 (2020), pp. 144–150. DOI: 10.1109/MRA.2019.2927372. URL: https://doi.org/10.1109/MRA.2019.2927372; Friederike Eyssel and Frank Hegel. "(S)he's got the look: Gender-stereotyping of social robots". In: *Journal of Applied Social Psychology* 42.9 (2012), pp. 2213–2230. DOI: 10.1111/j.1559-1816.2012.00937.x. URL: https://doi.org/10.1111/j.1559-1816.2012.00937.x; Gwen Bergner. "Black children, White preference: Brown v. Board, the doll tests, and the politics of self-esteem". In: *American Quarterly* 61.2 (2009), pp. 299–332. DOI: 10.1353/aq.0.0070. URL: https://doi.org/10.1353/aq.0.0070; Friederike Eyssel and Steve Loughnan. "'It don't matter if you're Black or White'? Effects of robot appearance and user prejudice on evaluations of a newly developed robot companion". In: *Lecture Notes in Computer Science* 8239 (2013), pp. 422–431. DOI: 10.1007/978-3-319-02675-6_42. URL: https://doi.org/10.1007/978-3-319-02675-6_42.

[52] Jasmin Bernotat, Friederike Eyssel, and Janik Sachse. "Shape It – The Influence of Robot Body Shape on Gender Perception in Robots". In: *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings*. Ed. by Abderrahmane Kheddar et al. Cham: Springer International Publishing, 2017, pp. 75–84. ISBN: 978-3-319-70022-9. DOI: 10.1007/978-3-319-70022-9_8. URL: https://doi.org/10.1007/978-3-319-70022-9_8; Eyssel and Hegel, "(S)he's got the look: Gender-stereotyping of social robots"; Jahna Otterbacher and Michael Talias. "S/He's Too Warm/Agentic!: The Influence of Gender on Uncanny Reactions to Robots". In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '17. Vienna, Austria: ACM, 2017, pp. 214–223. ISBN: 978-1-4503-4336-7. DOI: 10.1145/2909824.3020220. URL: http://doi.acm.org/10.1145/2909824.3020220; M. Siegel, C. Breazeal, and M. I. Norton. "Persuasive Robotics: The influence of robot gender on human behavior". In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. New York: IEEE, 2009, pp. 2563–2568. DOI: 10.1109/IROS.2009.5354116.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

stereotypes and prejudices present in society may then be transferred onto humanoid robots. Changing the features of a robot to achieve this racialization can be as easy as changing the accent of the text-to-speech engine. All changes that can be accomplished by only changing the software are relatively easy. Moreover, once implemented for one robot, they can be copied to any other robot.

5.2 Impact of racialisation

Back in 2012, Eyssel and Kuchenbrandt[53] showed that ascribing ethnicity to a robot changes how people interact with it. Participants in their study not only rated in-group robots more positively, but they also anthropomorphised them more.

Several years later and inspired by ideas from Robert Sparrow, Bartneck et al.[54] showed that robots are racialised and that this in turn impacts the behaviour towards them. This study extended the 'shooter bias' experiment[55] by including robotic stimuli. The original experiment used white and black people while the robot study was only able to adjust the colour of the robot because other racial features, such as facial features, could not be changed. Still, participants not only racialised the robots implicitly, but also explicitly. Moreover, this racialisation resulted in the participants changing their behavior by exhibiting greater levels of 'shooter bias' (i.e., increased tendency to mistakenly shoot unarmed Black vs. White targets).

This experiment was then further extended to include further shades of brown to the original experiment.[56] The previously observed 'shooter bias' disappeared, perhaps because introducing more diverse racialisations increased the complexity of the task making color a less relevant heuristic. It was not clear what race or ethnicity the brown robot represents and hence what stereotype apply. The brown robot could be perceived as South Asian, Hispanic, or South East Asian making it difficult to know what cultural stereotype people may apply to this specific robot as these groups have differing stereotypes associated with them.

At least two more empirical studies were published in recent years. Strait et al.[57] showed that comments made to YouTube videos of robots racialised as Black, White and Asian robots differ considerably. More specifically, Black or Asian robots are dehumanised more than White robots. K. Barfield[58] confirmed that people do exhibit a racial bias towards robots. Participants in her study exhibit discriminatory views not only towards black robots, but also towards rainbow colorised robots.

While the number of empirical studies on racism in HRI is still relatively small, it has triggered a

[53] Friederike Eyssel and Dieta Kuchenbrandt. "Social categorization of social robots: Anthropomorphism as a function of robot group membership". In: *British Journal of Social Psychology* 51.4 (2012), pp. 724–731. DOI: 10.1111/j.2044-8309.2011.02082.x. URL: https://doi.org/10.1111/j.2044-8309.2011.02082.x.

[54] Bartneck et al., "Robots and Racism".

[55] Correll et al., "The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals".

[56] Addison, Yogeeswaran, and Bartneck, "Robots Can Be More Than Black And White: Examining Racial Bias Towards Robots".

[57] Strait et al., "Robots Racialized in the Likeness of Marginalized Social Identities are Subject to Greater Dehumanization than those racialized as White".

[58] Jessica K. Barfield. "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*. UMAP '21. Utrecht, Netherlands: Association for Computing Machinery, 2021, pp. 109–114. ISBN: 9781450383677. DOI: 10.1145/3450614.3463411. URL: https://doi.org/10.1145/3450614.3463411.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

much larger discussion on the theoretical,[59] cultural[60] and ethical[61] foundations of bias in the context of humanoids. The possibility that people will ascribe human identities to robots and that this categorization of humanoid robots may impact how we evaluate and interact with robots has become an important emerging direction of research in HRI.

6 Future Directions Emerging from Considerations of the Existing Gaps and Implications of Bias in HRI

6.1 The Unknowns

The current chapter considers the extent to which people's tendency to show bias toward other humans on the basis of their group membership can be ascribed to how we perceive and express bias towards humanoid robots. Although robots are not living beings in a biological sense, we appear to carry over biases we have toward humans on our evaluation of robots based on its appearance, voice, or perceived national origin. Such research highlights how the human perceptions of robots can be clouded by the perceived identities ascribed to robots. Even if the robots are not designed to resemble a specific group, they might still be perceived in some cases as having the attributes of human groups which influence people's attitudes, beliefs, and behaviors toward the robot. With that said, as this area of research is in its infancy, it is important to acknowledge the limitations of current empirical knowledge on the topic and consider valuable directions for future work.

First, while the research described in the chapter thus far has been valuable in highlighting how human biases can be applied to humanoid robots, it is important to note that such work has been limited to a handful of national contexts, and even more so, these have been focused on WEIRD (western, educated, industrialised, rich, democratic) nations.[62] Therefore, it is important to replicate and expand such work to other contexts. For example, it is unlikely that shooter bias tendencies would emerge in cultural contexts where stereotypes between dark and danger are likely to be paired or associated. In fact, research among humans reveals that shooter biases are driven by cultural stereotypes that people have and not based primarily on personal self-reported biases; thus even African American participants can reveal shooter bias tendencies. Therefore, future work should more broadly examine varied forms of implicit biases toward technology that is ascribed a human group identity to examine its prevalence and limits. Such work can test, for example, if the perceived physical

[59] Hee Rin Lee et al. "Configuring Humans: What Roles Humans Play in HRI Research". In: *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '22. Sapporo, Hokkaido, Japan: IEEE Press, 2022, pp. 478–492. DOI: 10.5555/3523760.3523824; Joshua Skewes, David M. Amodio, and Johanna Seibt. "Social robotics and the modulation of social perception and bias". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1771 (2019), p. 20180037. DOI: 10.1098/rstb.2018.0037. URL: https://doi.org/10.1098/rstb.2018.0037; Jean-Christophe Giger et al. "Humanization of robots: Is it really such a good idea?" In: *Human Behavior and Emerging Technologies* 1.2 (2019), pp. 111–123. DOI: 10.1002/hbe2.147. URL: https://doi.org/10.1002/hbe2.147.

[60] M. John Lamola. "An ontic–ontological theory for ethics of designing social robots: a case of Black African women and humanoids". In: *Ethics and Information Technology* 23.2 (2021), pp. 119–126. ISSN: 1572-8439. DOI: 10.1007/s10676-020-09529-z. URL: https://doi.org/10.1007/s10676-020-09529-z; Stephen Cave and Kanta Dihal. "The Whiteness of AI". in: *Philosophy & Technology* 33.4 (2020), pp. 685–703. ISSN: 2210-5441. DOI: 10.1007/s13347-020-00415-6. URL: https://doi.org/10.1007/s13347-020-00415-6; Laura Schelenz. "Artificial Intelligence Between Oppression and Resistance: Black Feminist Perspectives on Emerging Technologies". In: *Artificial Intelligence and Its Discontents: Critiques from the Social Sciences and Humanities*. Ed. by Ariane Hanemaayer. Cham: Springer International Publishing, 2022, pp. 225–249. ISBN: 978-3-030-88615-8. DOI: 10.1007/978-3-030-88615-8_11. URL: https://doi.org/10.1007/978-3-030-88615-8_11.

[61] Massimiliano L. Cappuccio et al. "Can Robots Make us Better Humans?" In: *International Journal of Social Robotics* 13.1 (2021), pp. 7–22. ISSN: 1875-4805. DOI: 10.1007/s12369-020-00700-6. URL: https://doi.org/10.1007/s12369-020-00700-6; Robert Sparrow. "Robotics Has a Race Problem". In: *Science, Technology, & Human Values* 45.3 (2019), pp. 538–560. ISSN: 0162-2439. DOI: 10.1177/0162243919862862. URL: https://doi.org/10.1177/0162243919862862.

[62] Joseph Henrich, Steven J Heine, and Ara Norenzayan. "The weirdest people in the world?" In: *Behavioral and brain sciences* 33.2-3 (2010), pp. 61–83. DOI: 10.1017/S0140525X0999152X. URL: https://doi.org/10.1017/S0140525X0999152X.

attractiveness of a robot influences how it is treated in a way that humans evaluate other humans. Alternately, future work can examine if similar to the vast literature on obesity stigma, people show more negative evaluations of a robot that appears 'overweight'.

Second, as research thus far has only examined the influence of ascribing race or gender to humanoids in isolation, it would be valuable for future work to consider the influence of multiple ascribed identities simultaneously. As a humanoid robot can simultaneously be ascribed a race or national origin along with gender or social class, we may respond differently to humanoids based on an interaction of these characteristics. Therefore, it would be valuable for future work to examine the potential interactive or additive impact of intergroup biases that may emerge based on a robot's perceived gender, race, social status, age, or physical appearance. Extensive research on multiple identities has been conducted in the context of human-human relations (for a review, see Kang and Bodenhausen[63]). However, while intuitively one may expect an additive or interactive impact of biases emerging from a robot's perceived race, gender, and social class, recent research suggests such effects may not be so clear cut. At least in the context of human biases, people's implicit evaluations appear to be predominantly influenced by a single social category. For example, Connor et al.[64] measured implicit evaluations of human targets varying in their race, gender, social class, and age. Across multiple studies, Connor and colleagues found that target gender was the primary determinant of people's implicit evaluations with female targets evaluated more favorably than male targets. They similarly found a smaller but consistent finding that upper-class targets were evaluated more favorably than lower-class targets. However, implicit evaluations toward targets with multiple distinct social identities were largely influenced by evaluations toward a single dominant category. While there was some evidence of a compounded effect such that the most negative implicit evaluations were toward lower-class males and the most favorable evaluations were toward upper-class females, these effects were considerably smaller relative to the impact of a single dominant category (i.e., gender, and then social class, in their research).[65] Taken together, these findings suggest that responses to targets that are multiply categorizable are predominantly driven by a single dominant category, but less dominant social categories can exert a small additive effect in people's evaluations. However, future work should examine how people's attitudes, beliefs, and behaviors toward humanoids are influenced by the simultaneous ascription of multiple human identities to humanoid robots.

6.2 Societal Implications

While treating one group of people better or worse than another has been the source of much misery and conflict throughout history and in contemporary society, robots are unable to experience any pain or suffering, so why may it matter if robots are differentially treated based on their perceived human attributes? One potential implication of people's tendency to treat humanoids differently based on their ascribed identities is that it could reinforce existing stereotypes and prejudices in society. If users treat a brown robot much worse than a white robot, or if robots destined for roles of authority or that of role models (e.g., teachers, doctors)[66] were all light colored, while robots doing subservient roles in our house or society at large happened to be dark colored, it may reinforce stereotypic beliefs or prejudiced feelings toward dark colored robots. At least in the context of human to human relationships, (counter)stereotypic exposure can influence our biases toward them. For example, reading news stories about Black (vs. White) criminals increased shooter bias tendencies by making stereotypes more accessible.[67] Similarly, exposure to admired African Americans reduced

[63] Sonia K. Kang and Galen V. Bodenhausen. "Multiple Identities in Social Perception and Interaction: Challenges and Opportunities". In: *Annual Review of Psychology* 66.1 (2015). PMID: 25061671, pp. 547–574. DOI: 10.1146/annurev-psych-010814-015025. URL: https://doi.org/10.1146/annurev-psych-010814-015025.

[64] Paul Connor et al. "Intersectional implicit bias: Evidence for asymmetrically compounding bias and the predominance of target gender". In: *Journal of Personality and Social Psychology* 124.1 (2023), pp. 22–48. DOI: 10.1037/pspa0000314. URL: https://doi.org/10.1037/pspa0000314.

[65] Ibid.

[66] Fady Alnajjar et al. *Robots In Education*. Routledge, 2021.

[67] Correll et al., "The influence of stereotypes on decisions to shoot".

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

implicit prejudice toward African Americans[68] while exposure to female leaders could increase implicit associations between women and leadership.[69]

While the research discussed above suggests that how dark vs. light colored robots are depicted could have spillover effects by increasing or decreasing certain implicit associations between a certain group with specific characteristics or valence, it is unclear if this will have any meaningful effect on how humans groups are perceived. After all, as self-categorization theory[70] reveals, people can perceive themselves at varying levels of abstraction including at the abstract level as a human being. People may then perceive a robot as simply as a distinct non-human entity and any associations they are exposed may simply not carry over to human social groups. Therefore, it is not clear whether exposure to robots in varying roles would indeed matter or not. Future research should examine the influence of having robots in different roles and whether exposure to robots experiencing differential treatment would indeed influence how humans treat each other. For example, while it was presumed from decades of research that there was a clear and robust relationship between exposure to violent media and real-world aggression,[71] even warranting public statements by the American Psychological Association (APA), a re-evaluation of the data suggested there is a negligible relationship between violent media and aggressive behavior, even if there was a more robust relationship with desensitization.[72] While debate is bound to continue on the extent to which violent media exposure causes real world aggression (also see[73]), such debate illustrates two important points. First, it is important not to assume that exposure to robots in varied roles in society or a lack of representation of different colored would meaningfully impact any real-world outcomes for human groups in society. And second, the debate on violent media and aggression also illustrates the importance of replicating and establishing a cumulative body of evidence on a topic before drawing conclusions.

In a similar vein, robots can be programmed to agree or even to enjoy abusive behavior towards them. A female robot could, for example, be programmed to enjoy non-consensual sex, in the same way actors can be paid for acting out rape pornography. In many countries, depictions of the latter are illegal. The argument against it is, however, not based on an expectation that people engaging or consuming such interactions will themselves become rapists. Instead, Sparrow[74] pointed out that such negative behaviour towards robots is "morally problematic for what it expresses about women and about the character of the user." As highlighted in the above discussion around violent video games,[75] more work is needed to determine if similar issues and developments occur in the context of

[68] Nilanjana Dasgupta and Anthony G Greenwald. "On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals". In: *Journal of personality and social psychology* 81.5 (2001), pp. 800–814. DOI: 10.1037/0022-3514.81.5.800. URL: https://doi.org/10.1037/0022-3514.81.5.800.

[69] Dasgupta, "Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations".

[70] Turner et al., *Rediscovering the social group: A self-categorization theory*.

[71] Craig A Anderson et al. "Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review". In: *Psychological bulletin* 136.2 (2010), pp. 151–173. DOI: 10.1037/a0018251. URL: https://doi.org/10.1037/a0018251; Brad J Bushman and Craig A Anderson. "Media violence and the American public: Scientific facts versus media misinformation". In: *American Psychologist* 56.6-7 (2001), pp. 477–489. DOI: 10.1037/0003-066X.56.6-7.477. URL: https://doi.org/10.1037/0003-066X.56.6-7.477.

[72] Christopher J Ferguson, Allen Copenhaver, and Patrick Markey. "Reexamining the findings of the American Psychological Association's 2015 task force on violent media: A meta-analysis". In: *Perspectives on Psychological Science* 15.6 (2020), pp. 1423–1443. DOI: 10.1177/174569162092766. URL: https://doi.org/10.1177/174569162092766; Aaron Drummond, James D Sauer, and Christopher J Ferguson. "Do longitudinal studies support long-term relationships between aggressive game play and youth aggressive behaviour? A meta-analytic examination". In: *Royal Society open science* 7.7 (2020), p. 200373. DOI: 10.1098/rsos.200373. URL: https://doi.org/10.1098/rsos.200373; Christopher J Ferguson. "Aggressive video games research emerges from its replication crisis (sort of)". In: *Current opinion in psychology* 36 (2020), pp. 1–6. DOI: 10.1016/j.copsyc.2020.01.002. URL: https://doi.org/10.1016/j.copsyc.2020.01.002.

[73] Anna T Prescott, James D Sargent, and Jay G Hull. "Metaanalysis of the relationship between violent video game play and physical aggression over time". In: *Proceedings of the National Academy of Sciences* 115.40 (2018), pp. 9882–9888. DOI: 10.1073/pnas.1611617114. URL: https://doi.org/10.1073/pnas.1611617114.

[74] Robert Sparrow. "Robots, Rape, and Representation". In: *International Journal of Social Robotics* 9.4 (2017), pp. 465–477. ISSN: 1875-4805. DOI: 10.1007/s12369-017-0413-z. URL: https://doi.org/10.1007/s12369-017-0413-z.

[75] Simone Kühn et al. "Does playing violent video games cause aggression? A longitudinal intervention study". In: *Molecular Psychiatry* 24.8 (2019), pp. 1220–1234. ISSN: 1476-5578. DOI: 10.1038/s41380-018-0031-7. URL: https://doi.org/10.1038/s41380-018-0031-7; Christopher J. Ferguson. "Violent Video Games, Sexist Video Games, and the Law: Why Can't We Find Effects?" In: *Annual Review of Law and Social Science* 14.1 (2018), pp. 411–426. ISSN:

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

humanoid robots such that there are real-world implications that warrant legal and policy regulations to mitigate risks to society.

Assuming bias in HRI has negative implications for society at large, it raises the question of how to fix this problem. Should we entirely shelf the idea of humanoid robots and settle for less anthropomorphic robots that are devoid of membership in any group? This would be difficult, if not impossible, as robots in many roles will need a voice and be manufactured somewhere, leading people to ascribe gender and a national origin to the technology. Moreover, anthropomorphic robots may be better suited and more effective in certain roles (for a review, see[76]), so opting for less anthropomorphic robots to tackle a bias problem will come with a cost. How then do we solve a bias problem in HRI? Even within human-human intergroup relations, there are no easy answers for how to reduce bias (for reviews, see[77]) after more than 50 years of research. Moreover, it is unclear if the most effective strategies to reduce bias in the context of human-human relations (e.g., positive contact; for reviews, see[78]) could be effective in the context of HRI. Would it help then to rely on some kind of universal design or custom build humanoids to be relevant to the country of sale similar to the option of selecting an accent and language on voice assistant devices? More empirical work is needed to answer these questions.

6.3 Market implications

Robots are not the first product to be racialized. The most commonly studied racialization occurred with dolls. As a result, in 1980, the toy company Mattel offered its first Black Barbie doll to consumers. It still used the same mold as the other Barbie dolls and hence still had many Caucasian features, while changing the color of the plastic was relatively easy. Still, given the large number of Barbie dolls being sold every year, the costs for a dedicated mold to change the shape of Black dolls would certainly be affordable for Mattel. Nevertheless, it took Mattel another ten years before they started to change the mold. The effect that multicultural and multiethnic dolls have on the development of children remains a heavily discussed topic.[79] Even in more recent history the sales of Barbie dolls remains a hot topic in the media. In 2010 and 2015, shoppers reported that the Black version of a certain Barbie doll was sold cheaper than its Caucasian counterpart which resulted in a public outcry.[80] From a purely commercial perspective, it could be argued that selling less popular items at a discount made sense. But since Barbie dolls represent humans, selling a Barbie of a certain race at lower value could be considered problematic.

Another market implication to consider is whether robots should be created in ways to cater to the diversity of the population or whether it would be more suitable to design robots that are entirely non-humanoid (e.g., robots that look like a vacuum cleaner or a tree). Offering a diverse range of

1550-3585. DOI: 10.1146/annurev-lawsocsci-101317-031036. URL: https://doi.org/10.1146/annurev-lawsocsci-101317-031036; Christopher J. Ferguson and Sven Smith. "Examining homicides and suicides cross-nationally: Economic factors, guns and video games". In: *International Journal of Psychology* 56.5 (2021), pp. 812–823. ISSN: 0020-7594. DOI: 10.1002/ijop.12760. URL: https://doi.org/10.1002/ijop.12760.

[76] Zlotowski et al., "Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction".

[77] Elizabeth Levy Paluck et al. "Prejudice Reduction: Progress and Challenges". In: *Annual Review of Psychology* 72.1 (2021). PMID: 32928061, pp. 533–560. DOI: 10.1146/annurev-psych-071620-030619. URL: https://doi.org/10.1146/annurev-psych-071620-030619; Elizabeth Levy Paluck and Donald P. Green. "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice". In: *Annual Review of Psychology* 60.1 (2009). PMID: 18851685, pp. 339–367. DOI: 10.1146/annurev.psych.60.110707.163607. URL: https://doi.org/10.1146/annurev.psych.60.110707.163607.

[78] John F. Dovidio et al. "Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions". In: *Group Processes & Intergroup Relations* 20.5 (2017), pp. 606–620. DOI: 10.1177/1368430217712052. URL: https://doi.org/10.1177/1368430217712052; Ananthi Al Ramiah and Miles Hewstone. "Intergroup contact as a tool for reducing, resolving, and preventing intergroup conflict: Evidence, limitations, and potential". In: *American Psychologist* 68.7 (2013), pp. 527–542. DOI: 10.1037/a0032603. URL: https://doi.org/10.1037/a0032603.

[79] Bergner, "Black children, White preference: Brown v. Board, the doll tests, and the politics of self-esteem"; Ann DuCille. *Skin trade*. Harvard University Press, 1996.

[80] Katie Little. *Big retailers price black, white Barbies differently*. Newspaper Article. 2014. URL: https://www.cnbc.com/2014/12/09/wal-mart-toys-r-us-price-black-white-barbies-differently.html.

options can be suitable in some contexts. For example, Johnson and Johnson started to sell Ebon-Aid adhesive bandages in the 1990s with a darker skin tone than the only light flesh color options they had in the US since the 1920s.[81] These were immediately popular with darker skinned people, and since then, a wider ranger of colours are widely available.

Humanoid robots have not yet become a mass market product. They are predominately used in the context of research and we do not know if and when they might become a commercial success.[82] Still, robot manufactures have to address questions about the racialization of their robots. Even if they do not consciously design the robot to have racial characteristics, they might still be perceived as such. Developers will have to take control of their designs. Sparrow[83] suggested that the one way robot developers can avoid these challenges is to design a robot that cannot easily be racialized. This will require a deep understanding of how robots are racialised and what features contribute to it. One approach could be to give robots distinct non-human skin colors, such as bright blue or green. The first design of the Optimus robot from Tesla, for example, complimented its 90% white body with black shoulders and head. The first operational prototype, however, focused on the underlying technology and did not include any type of cover or "skin". This demonstrates that engineers typically focus on functionality and less on the appearance of robots.

6.4 Legal implications

The relationship of people to technology is increasingly considered in legal settings. Several technologies that are closely aligned to robots have become a topic of legal debates. The European Commission discussed giving personhood to AI systems in 2017[84] which lead to a vivid discussion[85] amongst politicians and experts.[86] The Sophia humanoid robot was given citizenship in Saudi Arabia in 2018, although it has been pointed out that she will be condemned to a "lifeless career in marketing".[87] Giving robots the legal status of a person or even citizenship might still be better suited for future generations since current robots clearly do not show any signs of personhood or consciousness. In particular the latter seems to be a prerequisite for a legal status.[88] But it could be argued that robots could receive some legal standing. An often discussed approach would be to give them rights similar to animals.[89] While animals are still treated like property and owners have every right to kill them, they cannot be cruel to them. Thousands of animals are killed every day for the production of food. Male chicks even get killed purely to optimise profits. Calverley[90] pointed out that animal rights might be a stepping stone for robot rights, although a distinction should be made when contrasting living being like animals that feel pain from an inanimate object such as a robot.

[81] Sebastien Malo. *The Story of the Black Band-Aid*. Newspaper Article. 2013. URL: https://www.theatlantic.com/health/archive/2013/06/the-story-of-the-black-band-aid/276542/.

[82] C. Bartneck. "Why do all social robots fail in the market?" In: *The Human-Robot Interaction Podcast* 15 (2020). DOI: 10.17605/OSF.IO/7KFRZ. URL: https://www.human-robot-interaction.org/2020/10/19/why-do-all-social-robots-fail-in-the-market/.

[83] Sparrow, "Robotics Has a Race Problem".

[84] https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html?redirect

[85] http://www.robotics-openletter.eu

[86] Pin Lean Lau. "The Extension of Legal Personhood in Artificial Intelligence Dossier sobre Inteligencia Artificial, Robotica e Internet de las Cosas". In: *Revista de Bioetica y Derecho* 46 (2019), pp. 47–66. URL: https://heinonline.org/HOL/P?h=hein.journals/rebiod46&i=55https://heinonline.org/HOL/PrintRequest?handle=hein.journals/rebiod46&collection=journals&div=6&id=55&print=section&sction=6.

[87] Emily Reynolds. *The agony of Sophia, the world's first robot citizen condemned to a lifeless career in marketing*. Newspaper Article. 2018. URL: https://www.wired.co.uk/article/sophia-robot-citizen-womens-rights-detriot-become-human-hanson-robotics.

[88] Maartje M.A. de Graaf, Frank A. Hindriks, and Koen V. Hindriks. "Who Wants to Grant Robots Rights?" In: *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '21 Companion. Boulder, CO, USA: Association for Computing Machinery, 2021, 38–46. ISBN: 9781450382908. DOI: 10.1145/3434074.3446911. URL: https://doi.org/10.1145/3434074.3446911.

[89] Kate Darling. *The new breed: what our history with animals reveals about our future with robots*. Henry Holt and Company, 2021. ISBN: 9781250296108. URL: https://www.worldcat.org/title/1224045568.

[90] D. J. Calverley. "Android science and animal rights, does an analogy exist?" In: *Connection Science* 18.4 (2006), pp. 403 –417. DOI: 10.1080/09540090600879711. URL: https://doi.org/10.1080/09540090600879711.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

An important legal discussion in HRI with human lives at stake are with the use of Autonomous Vehicles (AV). AVs have been involved in many road accidents that lead to damages, injuries, and even deaths. For example, in 2019, a Tesla car with an active auto pilot ran a red light and killed Gilberto Alcazar Lopez and Maria Guadalupe Nieves-Lopez.[91] It is arguably the first case involving felony charges to result from a fatal crash involving an automated driving system. The court case only started in 2022 and there is still no clear cause of action for this case. Who will be responsible for this tragic accident: Tesla, the driver Kevin Riad, or both? On what basis will the judges decide on the responsibilities? Even more important, why has the US legal system not already provided the necessary guidance for the judges? The UK took first steps to regulate strict liability for AV accidents by introducing a road map that will make the manufactures responsible for AV accidents when they drive in autopilot.[92] All countries will have to come up with a legal framework since AVs are being sold worldwide.

While a legal framework will have to be developed for certain types of robots, such as AVs, it is less obvious if, how, and when legislation will have to be introduced that would apply to social robots. We hope that the book at hand will provide a good foundation for the necessary discussions. In the same way that Mattel cannot be forced to sell brown dolls, a robot manufacturer could also not be easily forced to offer more variety (and it is unclear how critical this would be). Laws and policies are often necessary to systematically introduce change. However, legal changes are subject to backlash and may also have unintended negative consequences. In the context of humanoid robots that are ascribed human identities, it is unclear if any change is required, and if so, what regulation would be needed. For example, if it is important to have a variety of humanoid robots, the market could potentially self regulate, making it financially attractive for robot developers to develop a greater variety of robots. Similar to how the skin colors of Emojis have become more diverse after new skin tones were introduced, there may be greater demand for a wider range of robot colors over time. However, another option would simply be to only create non-human colored robots for the public. If research suggests that exposure to racialized robots in varying roles impacts human prejudice and discrimination, then it would be important for policy makers to try and ensure that different colored robots do not systematically occupy specific roles. However, without research directly examining the influence that robot representation in varied roles for human biases, the legal and ethical implications remain unclear. While the current chapter does not offer guidance for the best path forward, we hope this chapter provides a starting point for future exploration in this emerging area of study within HRI.

7 Summary and Conclusions

The emergence of humanoid robots has raised an interest in understanding whether we ascribe human social identities to such technology. Such humanoids will inevitably have to take on a physical form, have a voice, and be made somewhere, so the question then is how are we influenced by the social identities we ascribe to these novel technologies? Recent empirical research in HRI suggests that we knowingly and unknowingly can ascribe human social group identities to robots based on appearance, voice, and other features. And more importantly, our attitudes, beliefs, and behaviors can be influenced by projecting human social group affiliations to technology. In the current chapter, we explore these complex questions and consider the extensive psychological literature on human intergroup relations to identify gaps and valuable directions for understanding the path forward for research on humanoids. While one simple solution for the challenges highlighted above would be to develop functional robots that do not resemble humans in any way, this may have negative implications as anthropomorphized robots may better connect with human users (e.g., companion robots). So as long as we have humanoid

[91] Jonathan M. Gitlin. *Manslaughter charges follow Tesla driver's Autopilot red light run.* Journal Article. 2022. URL: https://arstechnica.com/cars/2022/01/manslaughter-charges-follow-tesla-drivers-autopilot-red-light-run/.

[92] Rebecca Bellan. *Human drivers in UK won't be liable for accidents when vehicle is self-driving.* Newspaper Article. 2022. URL: https://techcrunch.com/2022/08/21/human-drivers-in-uk-wont-be-liable-for-accidents-when-vehicle-is-self-driving/.

Yogeeswaran, K., & Bartneck, C. (2024). Racialization and Bias toward Humanoids. In W. Barfield, Y.-H. Weng, & U. Pagallo (Eds.), *The Cambridge Handbook of the Law, Policy, and Regulation for Human–Robot Interaction* (pp. 556–572). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009386708.036

robots, people may associate the robot with a group membership through its voice, appearance, place of origin, etc. So then it is important to understand how such ascription of human identities to humanoids impact our interaction with the technology, whether this has any impact on human-human relations, and if so, what can be done to mitigate such issues. The current chapter merely plants a seed for researchers in this nascent multidisciplinary area of study.