Sci Eng Ethics (2018) 24:1521–1536
https://doi.org/10.1007/s11948-017-9975-2

ORIGINAL PAPER

The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity

Ayanna Howard[2] · Jason Borenstein[1]

Received: 3 July 2017 / Accepted: 8 September 2017 / Published online: 21 September 2017

Abstract Recently, there has been an upsurge of attention focused on bias and its impact on specialized artificial intelligence (AI) applications. Allegations of racism and sexism have permeated the conversation as stories surface about search engines delivering job postings for well-paying technical jobs to men and not women, or providing arrest mugshots when keywords such as "black teenagers" are entered. Learning algorithms are evolving; they are often created from parsing through large datasets of online information while having truth labels bestowed on them by crowd-sourced masses. These specialized AI algorithms have been liberated from the minds of researchers and startups, and released onto the public. Yet intelligent though they may be, these algorithms maintain some of the same biases that permeate society. They find patterns within datasets that reflect implicit biases and, in so doing, emphasize and reinforce these biases as global truth. This paper describes specific examples of how bias has infused itself into current AI and robotic systems, and how it may affect the future design of such systems. More specifically, we draw attention to how bias may affect the functioning of (1) a robot peacekeeper, (2) a self-driving car, and (3) a medical robot. We conclude with an overview of measures that could be taken to mitigate or halt bias from permeating robotic technology.

✉ Jason Borenstein
borenstein@gatech.edu

Ayanna Howard
ayanna.howard@ece.gatech.edu

[1] School of Public Policy, Georgia Institute of Technology, 685 Cherry Street, Atlanta, GA 30332-0345, USA

[2] School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Keywords Artificial intelligence · Implicit bias · Design ethics · Professional ethics · Robot ethics

Introduction

Bias is a feature of human life that is intertwined, or used interchangeably, with many different names and labels—stereotypes, prejudice, implicit or subconsciously held beliefs, or close-mindedness. A standard view is that bias is a negative phenomenon which must be overcome. However, classifying the natural tendency to have biased attitudes or beliefs as ethically right or wrong is not the major issue for discussion here. Rather, we are focused herein on when and how bias influences the decision-making process; when an individual makes a biased decision or when bias prevents an objective consideration of an issue or situation, ethical problems may arise.

Bias continues to be a large and pervasive problem that has countless, often detrimental, impacts on human well-being. Its effects can stem from interfering with employment decisions to influencing the quality of education or healthcare that an individual or group receives. In the digital age, bias has often been encoded in and manifests itself through machine learning algorithms. This certainly has had far-reaching consequences, including search engines that stigmatize women or automated judicial systems that discriminate against racial minorities.

As these algorithms evolve into advanced artificially-intelligent agents, intertwining with our physical world, the negative ramifications of bias only increase. It is therefore an opportune time to begin addressing how bias can influence the design and behavior of these embodied intelligent decision-making agents, otherwise known as robots. As roboticists or other designers intentionally or unintentionally make their biases manifest in the technologies they create in the future, a necessary goal is to more effectively mitigate the resulting ethical problems. Thus, our main aims in this paper are two-fold: first, we seek to describe how bias could influence the designs of next generation robots; second, we seek to identify strategies that may mitigate or prevent bias from negatively affecting those who interact with future robotic systems.

What is Bias?

Many different types of bias can interfere with decision-making processes. For instance, Chavalarias and Ioannidis (2010) articulate a taxonomy of 235 biases that can arise when conducting research; they are largely referring to mental processes or behaviors that can contaminate research projects. Bias is also mentioned with respect to performance problems directly correlated with class-imbalance in datasets used to prove (or disprove) a hypothesis or to train algorithms (Kotsiantis et al. 2006; Chawla et al. 2002).

Not all forms of bias manifest themselves in unfair or otherwise unethical acts. Often though, bias refers to the unfair beliefs or behaviors that one directs toward a particular individual or group, which is more in line with the sense of the term that we seek to discuss here. What normally underlies the ethically problematic nature of this form of bias is thinking about or treating another person differently based on perceived characteristics of the individual. It is frequently tied into preconceived notions about a person’s gender, race, age, or sexual orientation. Bias can manifest itself in both “positive” (favoritism) and “negative” (unjust discrimination) ways. For our purposes here, we will largely focus on the latter, pejorative sense of the term.

Examples of bias range from the everyday judgments and assumptions people make about one another (e.g., I do not trust someone because of that person’s strange outfit) to more systemic, widespread practices (e.g., one group of people receives a higher loan interest rate than another due to differences in racial or ethnic background). Bias can become entrenched and have detrimental effects even relatively early in life. For example, Bian et al. (2017) claim that girls even as young as six may fall prey to the stereotype that they are not as intelligent as boys their own age.

Bias can be particularly troubling because we as humans may not be consciously aware of it, a phenomenon referred to as “implicit bias”. Brownstein (2016) defines implicit bias as “a term of art referring to relatively unconscious and relatively automatic features of prejudiced judgment and social behavior.” Implicit bias not only sneaks itself into our day-to-day lives, but it can also contaminate professional decision-making processes. For example, a study by Green et al. (2007) indicates that a physician’s implicit biases about different racial groups could interfere with treatment decisions. Moreover, Hall et al. (2015) reviewed 15 health-related studies and concluded that implicit bias was significantly related to patient–provider interactions, treatment decisions, and patient health outcomes.

A possible outgrowth of bias is the problem of “stereotype threat” whereby the fear that others might believe something negative about you, oftentimes connected to concerns about your gender or racial identity, can contribute to the belief becoming a self-fulfilling prophecy (Steele 2010). For example, providing subtle or direct cues to women that they are not as good as a man at math can contribute to them performing more poorly on math problems than their male counterparts (Spencer et al. 1999).

The scope of the effects of bias, in terms of stereotypes or other discriminatory behaviors, on vulnerable populations (e.g., disabled individuals) can be especially devastating. Bias can be particularly worrisome if it seeps unnoticed into the realm of engineering and design. For example, even though there may be no deliberate intent to cause harm, creating sidewalks without curbs or buildings without elevators can severely disadvantage individuals with movement impairments or disabilities. It is noteworthy that some measures are already in place that seek to counteract the influence of bias on engineering projects, including implementing principles of universal design (Harpur 2013).

While the line between the two can sometimes be murky, the distinction between bias and treating others differently for “good” reasons is important. For example, it is arguably justifiable to charge teenage drivers more for car insurance even though in some sense, they are being discriminated against as a group. Of course, what counts as a “good” reason can be subject to extensive debate; it is an issue that philosophers and others have explored for quite some time (e.g., see McHugh and Way 2016), and one we do not seek to resolve here. Yet what may help distinguish bias from a defensible difference in treatment is in the latter case, it should be consciously acknowledged that an individual or group is going to be treated differently. And in such cases, proposed justifications should be vetted by the involved stakeholders—from the decision-makers to those individuals who are directly impacted by a relevant decision. Moreover, logic and evidence should be transparently presented when seeking to make the case that treating an individual or group differently is appropriate.

Bias in the Realm of Artificial Intelligence (AI)

There has been a recent upsurge of attention focused on bias in the realm of artificial intelligence (AI). While the term AI has many varied definitions, for the purpose of this paper, we use the generalized understanding that AI refers to the algorithms and machinery that learn and then act on that learning (Bogost 2017). Thus, AI, in this regard, represents the mechanism that powers the decision-making process of a computing agent—whether that agent is embodied in hardware and paired with a robot body or is embodied in software such as a chatbot executed on a computing platform (Nwana 1996). In this section, we discuss the realm of AI as it relates to the virtual/computing environment. Allegations of racism and sexism have permeated the conversation about AI as stories surface about search engines delivering job postings for well-paying technical jobs to men and not women (Carpenter 2015), or providing arrest mugshots when keywords such as "black teenagers" are entered (Guarino 2016).

AI algorithms are evolving; they are often created from parsing through large datasets of online information while having truth labels bestowed on them by crowdsourced masses. These specialized AI algorithms have been liberated from the minds of researchers and companies, and released onto the public. These algorithms may be learning and acting on that learning, but they maintain some of the same biases that permeate our society. They find patterns within datasets that reflect our own implicit biases and, in so doing, emphasize and reinforce these biases as global truth. Below we will offer several examples that illustrate how bias, in the sense mentioned previously, has infused itself into the realm of AI.

Face Recognition Applications and AI

In 2010, the media reported a number of cases in which facial recognition algorithms had difficulty identifying non-Caucasian faces. A young female, of Taiwanese decent, blogged about her camera labelling her as "blinking," which went viral as did a video in which black users complained of their webcam not being able to detect their faces (Rose 2010). One could argue that these algorithms were at the early stages of AI, and these types of mistakes were the underpinnings of new technologies being introduced into society. On the other hand, one could make the case that these devices should have been more thoroughly tested with a broader base of users.

A few years later, Google made a faux pas in labeling black people as “gorillas” in its new photos application for iOS and Android, an application with the goal of becoming an intelligent digital assistant (Pulliam-Moore 2015). In 2016, the first international beauty contest took place in which AI algorithms were used to identify the most attractive contestants from a set of roughly 6000 entrants from over 100 countries (Levin 2016). Alas, race seemed to play a larger role than anticipated in the algorithms’ thinking process; of the 44 winners, 36 were white. In another case, an automated passport application system could not detect that an applicant of Asian descent had his eyes open for a passport picture; after each attempt at a photograph, the system indicated that the applicant had his eyes closed (Griffiths 2016).

So, how did such biases get introduced into an AI agent that, theoretically, should be less judgmental than a human? After all, AI algorithms typically develop their decision-making capabilities by learning patterns from massive amounts of data. For facial recognition, these algorithms would have been fed troves of images with labels mined from a host of different “experts”. Of course, in these cases, bias can creep into the system through many different avenues—from human bias present in the selection of the experts, deciding from where to secure the images, or even in determining how to define the desired output set.

Voice Recognition Applications and AI

There has long been gender-bias connected to voice recognition systems and their differences in performance when comparing the voices of men and women (e.g., see Larson 2016). According to a study by Rodger and Pendharkar (2004), medical voice-dictation software was found to more accurately recognize voice input from a man versus a woman. In 2011, several carmakers acknowledged that integrated speech-recognition technology was more difficult for women to use than men when trying to get their vehicles to operate properly (Carty 2011). A 2016 study claims to have identified gender bias in Google’s speech recognition software (Tatman 2016). The study indicates that queries from male voices were more consistently understood by the software than those from women. Although Google uses AI to improve the performance of its speech recognition algorithms, biases still get introduced—this ranges from the corpus of annotated speech that is used to train its systems to the engineers that select the corpus to use in the first place. This is not necessarily to say that the engineers and others deliberately sought to encode bias into their algorithms; rather, it highlights and reiterates the point that bias often works its way unnoticed into technological artifacts.

Search Engine Applications and AI

Search engines and their associated AI algorithms are well-recognized offenders in the computing world when it comes to perpetuating social biases and stereotypes. Search engines have become one of our most trusted sources of information and, in many ways, have become arbitrators of truth. Given how much trust is placed in the technology, designers and coders carry with them significant ethical responsibilities for their creations. Search-engine technology is not “value-neutral”; instead it has built-in features in its design that tends to favor some “values” over others. The values embedded in search-engines can introduce bias, which as a by-product directly shapes our perceptions and our opinions based on what information we believe is available or true. For example, over the last few years, many high-profile “mistakes” have caught the public’s eye; they bring to light the social bias inherent in search engines (Otterbacher 2016). For instance, a study by Datta et al. (2015) claims that Google displays far fewer ads for high-paying executive jobs to women than to men.

In 2016, public outrage was sparked when it was discovered that searching for “three white teenagers” on Google image search resulted mostly in images of happy people whereas the search phrase “three black teenagers” offered an array of mug shots (Guarino 2016). A study by Kay et al. (2015) supports the claim that search engines reinforce social stereotypes. The researchers compared the gender distributions of search results associated with a given profession. They found that search results for “doctors” retrieved significantly more images of men whereas a search for “nurses” provided significantly more images of women.

Of course, some may argue that search engine algorithms are not inherently biased and are value neutral—i.e., they are just reflecting that, in certain situations, there are more instances of one categorization versus another present in our society. For instance, a search engine may associate more female images with the term “nurse” because, in the U.S., approximately 90% of nurses are female (Muench et al. 2015). Yet to counter this point, in the U.S., 34% of the active physician workforce in 2015 and 46% of physicians-in-training are female (AAMC 2016). Therefore, the rationale for search engines historically delivering significantly more images of men when entering the term “doctor” holds little weight. The physician example lends support to the view that AI algorithms are inheriting human societal biases.

As search engines do not use a set of criteria that is transparent to the user in generating their results, search engines easily can reflect the biases inherent in society, in their developers, and in their users. Search results may even vary depending on how much of a user’s profile is available to the algorithm. Yet corrective measures could be put in place during the training process for the AI that underlies these search engines. For instance, Chayes (2017) claims that “with careful algorithm design, computers can be fairer than typical human decision-makers, despite the biased training data.” This could help ensure that historical biases do not overshadow changing outlooks on social equity.

The Justice System and AI

Biases in the realm of AI create some immediate concerns when such systems, which one would hope are designed to be as fair and objective as possible, reinforce prejudicial thought processes. Concerns have been repeatedly expressed about how biased interpretations of big data might unjustly influence governmental or other policies (e.g., Hardt 2014). For example, algorithms are currently being used by the US justice system to predict whether a criminal is likely to perform future violent acts, which can affect decisions about sentencing and parole. Arguably, some of the relevant computing systems unfairly discriminate against African Americans by at times ranking them as being more dangerous than their counterparts when the reverse may be true (Angwin et al. 2016).

Along related lines, Robinson and Koepke discuss the practice of "predictive policing" whereby police departments rely on data from past criminal activities to guide future practices. They claim that "predictive systems that rely on historical crime data risk fueling a cycle of distorted enforcement" (Robinson and Koepke 2016). Predictive policing may seem to be an effective measure to prevent criminal activity. Yet the biases present in the algorithms can have long-reaching consequences for specific individuals and groups of people. This goes well-beyond the injustices felt when an individual from a certain gender or race is categorized as blinking or the inconvenience experienced when a car stereo does not recognize a voice command to change the radio station. The stakes are much higher in this circumstance. For example, the use of biased information could entail an extended and undeserved period of incarceration, which unjustly affects those who are arrested and possibly ruins the lives of their families.

Although it is difficult to quantify systematically the degree to which it occurs, bias in the realm of AI persists as the above examples should make clear. Furthermore, the essential, integral links between AI and robotics entail that bias can and will pervade the latter realm. What follows is that systematic analysis is needed to identify: (1) how bias might influence the design and use of robots and (2) which mechanisms can mitigate the effects of such bias.

Bias and Robotics

To further expand the discussion of bias in the realm of AI as it relates to the virtual/computing environment, in this section, we focus on robotics, which we classify as computing agents embodied in hardware and paired with a mechanical body. Fully-autonomous robots are still in the infancy stage with respect to their development and inclusion in society. However, for the purposes here, we examine robots based on predictions of how they will develop in the future; we leave open the question of whether non-embodied computing entities, like chatbots, fit within the definition of a robot.

Over the last several years, many organizations have formed and reports released that focus on AI and its impact on society. These efforts range from IEEE and its *Guidelines for Ethically Aligned Design for AI and Autonomous Systems* (IEEE 2017) to the recently announced $27 million Ethics and Governance of AI fund (MIT Media Lab 2017) and Stanford's 100-year study on AI (Stone et al. 2016). Other relevant initiatives such as a report on *Artificial Intelligence, Automation and The Economy* by former President Obama's administration (EOP 2016) and the global conglomeration of companies that make up the Partnership on AI (Finley 2016) also mention the problematic nature of bias. Within the aforementioned items, there tends to be a brief mention of robots and/or the physical realization of other AI agents. Yet robots are currently being designed to be used in numerous environments and contexts. To illustrate the potential scope and magnitude of the problem of bias in the realm of robotics, we will focus on three significant robot use cases: (1) the robot peacekeeper, (2) the self-driving car, and (3) the medical robot.

The Robot Peacekeeper

As mentioned previously, AI systems are currently being used by the U.S. justice system and arguably, these systems unfairly discriminate against minorities, especially African Americans, including when it comes to decisions about sentencing and parole. Robots are also making their way into the realm of law enforcement. In 2016, it was reported that a police robot had a direct role in taking the life of a man in Dallas (McFarland 2016). A key facet of the case is the man happened to be African-American. Given the current tensions arising from police shootings of African American men from Ferguson to Baton Rouge, it is disconcerting that robot peacekeepers, including police and military robots, will, at some point, be given increased freedom to decide whether to take a human life, especially if problems related to bias have not been resolved.

In the future, robots can, and will, be used in a variety of security and policing situations. In fact, North Dakota passed a law in 2015 that allows the police to use armed drones (Wagner 2015). As a robot police officer is tasked with patrolling busy metropolitan streets, it will need to process an enormous amount of data associated with monitoring the environment and the sheer number of people transitioning through that environment. Presumably, the robot will not be able to monitor each person's movements and actions in a large crowd with equal diligence. As such, the robot may need to use algorithms, such as a selective attention mechanism, to efficiently manage its computing resources and enable it to quickly concentrate on specific objects or events of interest. Presumably, the overarching goal would be the optimal "protection" of society. Developers may decide that a design to achieve this capability is by profiling individuals or groups in a similar way in which human police officers do. While "predictive policing" techniques could enable a robot to focus its attention, this design pathway opens the door to bias.

The Self-Driving Car

The promise of self-driving cars has drawn global attention. Expectations are high, but it is hard to determine whether the lofty stated goals, such as the technology saving millions of lives, are achievable (Freeman 2016). Over the next decade, car manufacturers ranging from Ford and Toyota to younger companies such as Tesla and NuTonomy expect to build cars that have a form of autonomy in the range of Level 4 or 5 (Hackett 2016). At Level 5, the system is supposed to operate in a manner that is equivalent to a human driver; it can go anyplace where it is legal to drive and make its own decisions during this process (SAE International 2016). Level 4 differs in that the system can drive and make decisions autonomously but not in all situations or environments. Unfortunately, all driving involves risk, and that risk can be further impacted by the scourge of bias.

Driverless cars will need to consider a range of alternatives when faced with a situation in which a crash is inevitable. For the foreseeable future, this may be due to the dynamic and chaotic nature of humans operating in the same space as a self-driving vehicle. "Trolley problem" types of decisions will involve making choices in terms of who gets put in harm's way (e.g., the driver, pedestrians, or even neighboring cyclists). In fact, MIT's Moral Machine website solicits input on these scenarios through crowdsourcing, basically trying to answer the question of who a driverless car should kill.[1] Yet a proposed solution to the "Trolley problem" raises the seemingly unavoidable problem of adding subjectivity to the ranking of the value of a life.

The design of driverless cars should be put under rigorous ethical scrutiny. The U.S. government recently released a checklist for self-driving cars which states that ethical judgments and decisions should be made consciously and intentionally (Kang 2016). Algorithms for resolving these conflict situations must be transparent and disclosed to the U.S. Department of Transportation (DOT) through the National Highway Traffic Safety Administration (NHTSA). However, these disclosure efforts do not seem to directly address the issue of bias and how its manifestations may arise in a self-driving car's decision-making process. For example, how is the value of one life going to be weighed over another? Reviewing the history of the differential treatment racial or other groups have received, including by the criminal justice system (ACLU 2014), reveals the vast and pervasive impact of implicit bias. Clarity in terms of recognizing the ethical problematic effects of bias typically happens retrospectively.

The Medical Robot

A vast array of robots assist in healthcare, including in surgery such as the da Vinci Surgical System or the TUG robots that deliver supplies and medication in hospitals. In 2016, a robotic startup that uses medical drones to deliver medicine and blood to areas in Rwanda launched a similar program to deliver supplies in rural areas of the US (Toor 2016). The push to develop healthcare robots will continue to gain momentum as their sophistication increases. Using robots to promote this type of societal good might be a noble cause. However, it is well known that humans, including nurses and doctors (Kane 2010), struggle with ethical decisions, whether it involves withdrawing life-sustaining therapy to disclosing an HIV-positive diagnosis to patients or their loved ones.

For a medical robot, the ethical framework for making such decisions must be enabled, whether with learning algorithms or through direct programming (Borenstein et al. 2017b). Assuming that humans are unable to fully remove bias from their decision-making process about ethical issues, it seems to follow that bias will be a strong influence on the decisions made by a medical robot. It is a natural conclusion considering that robots are programmed by fallible human beings. To illustrate the issue more tangibly, should a medical drone decide not to deliver supplies to an area with known terrorists? Or in a triage situation, should a medical

[1] Refer to http://moralmachine.mit.edu/ (accessed July 3, 2017).

robot provide assistance to a child first or an older adult (the latter with presumably a shorter life expectancy)? Just as human medical professionals have had, and will continue, to address such issues where reasonable people might disagree, medical robots will have to grapple with such ethically fraught issues as well.

Considerations with respect to bias in the robotics domain is at its infancy. Yet with the large upsurge in robotic investments, robotic startups and clear advances in the technology, bias, just as in the AI realm, will need to be addressed in the immediate future. As such, in the following sections, we delineate potential approaches for mitigating the biases of roboticists and others as well as strategies to consider when developing these robotic systems.

The Difficulty of Eradicating Bias

As previously stated, bias tends to effect human behaviors based on attitudes or stereotypes held against a particular individual or group. Oftentimes, it impacts the decision-making process in an unconscious manner (i.e., implicit bias). A key difficulty with eliminating bias, implicit or otherwise, is that it, and its associated attitudes about others, develop over a lifetime. They form based on our constant exposure, starting from a very young age, to both direct and indirect messages about others. For example, young students, when first learning about famous scientists in school, typically hear about idols such as Albert Einstein and Thomas Edison. A bias inadvertently takes root, namely that in order to be a scientist, you have to be male. This bias continues to strengthen with the child's exposure to non-female scientists—on television, in books, and in movies—eventually perpetuating into a firmly entrenched belief that may impact decisions as an adult. Along these lines, studies indicate that bias impacts perceptions of who is more qualified for a job in science (OSTP 2016, 14). In short, implicit bias can lead to differential, and often ethically problematic, treatment of others.

Another related difficulty with eliminating bias is that it may have evolved as a protective mechanism to enhance the decision-making process, especially during high-risk scenarios. As Gendler states (2011), "classifying objects into groups allows us to proceed effectively in an environment teeming with overwhelming detail." When faced with uncertainty about the world, biases may reactively save us from making detrimental mistakes. As Johnson et al. explain (2013), bias may sometimes cause us to mistakenly think, for example, that a stick is a snake but doing so helps us err on the side of caution when danger may be present.

One well-established instrument to evaluate bias is the Implicit Association Test (IAT); it measures attitudes and beliefs that people may be unwilling or unable to report.[2] Results from the IAT commonly reveal that human beings have their share of biases. What compounds the issue is that even when people are encouraged to recognize their implicit biases, behavioral change may unfortunately not follow. For example, workplace interventions designed to lessen the impact of bias have had mixed success (OSTP 2016, 14–15). Bias may be a feature of human life that cannot be completely eliminated, at least not at the present time. Yet one would hope that

[2] For more information refer to https://implicit.harvard.edu/implicit/ (accessed July 3, 2017).

systematic initiatives are implemented to mitigate its detrimental effects, and in this context, especially when it comes to robotics.

The Professional Responsibilities of Roboticists and Possible Solutions

In January 2016, a report released by the Federal Trade Commission emphasized that algorithms based on big data sets may "reproduce existing patterns of discrimination, inherit the prejudice of prior decision-makers, or simply reflect the widespread biases that persist in society" (FTC 2016). To deal with these issues, some believe that it remains the responsibility of individuals (i.e., humans) to scrutinize the behaviors of robots to ensure they are operating in an unbiased manner. In this way, decisions would still remain under the jurisdiction of an individual, organization, or other collective entity, and thus fault can be assigned. Yet relying only on humans to identify bias, given its unconscious facets, is not a suitable solution to fully address the issue. Eventually the collective nature of these implicit biases could lead to a systematic failure, perhaps leading to some form of harm to humans, including injury to pride, damage to trust, or even an intrusion upon one's liberty or pursuit of happiness. Without systematic and thorough ethical reflection, decisions may result that exclude certain demographics, perhaps based on factors such as age, race, gender, sexual orientation, or disability.

As Brownstein (2016) states, "One can ask whether agents are responsible for their implicit attitudes as such, that is, or whether agents are responsible for the effects of their implicit attitudes on their judgments and behavior." Holroyd (2012) argues that even though implicit bias can be difficult to identify and correct, one should be held morally responsible to the extent that it influences judgment or behavior. Along these lines, one cannot necessarily expect roboticists to change their attitudes about the world; yet they can in principle be held accountable for the effects of such attitudes when they are encoded in their designs. Thus, what approaches can the community employ to mitigate the potential biases of roboticists? And what can roboticists do to become more aware of bias when designing their robot platforms of the future?

Professional codes of ethics, including from organizations such as IEEE and the Association for Computing Machinery (ACM), normally offer general, well-intended guidance; yet insights from codes of ethics can be difficult to operationalize for a contentious ethical problem. As with many issues at the intersection of ethics and emerging technology, it may require individual practitioners and communities to go beyond the tenets of a code (Borenstein et al. 2017a). IEEE (2017) seems to recognize that it needs to go beyond "the codes" with its *Ethically Aligned Design* series of reports, which seek to grapple with ethically contentious issues emerging out of AI and robotics. IEEE's initiative could lead to the promulgation of technical standards, perhaps ones that have a direct or indirect connection to the issue of bias.[3]

[3] For example, see IEEE PROJECT: P7003—Algorithmic Bias Considerations, https://standards.ieee.org/develop/project/7003.html (accessed August 24, 2017).

Roboticists and other designers could seek to implement strategies that encourage participation from the public in the design process; it is especially important to involve segments of the population who are most likely to be directly affected by the technology. This is crucial because the robotics community should not assume that its design choices will fully map onto the public's preferences. For example, Šabanović et al. (2015) describe the implementation of a participatory design strategy involving older adults suffering from depression.

Community involvement can be a prerequisite for success. Thus, prior to the deployment of a new robot platform, the community should demand evidence that the relevant stakeholders were involved in the design process. For example, if the purchase of a robot peacekeeper is being considered, as part of the due-diligence contracting process, governing bodies should require proof that input was secured from citizens of large urban cities. Transparently sharing data from the procurement process with the public can also build perceptions of fairness.

Another recommendation is to encourage the formation of multidisciplinary teams. Those trained to recognize and mitigate implicit bias should be an integral part of the design process. The value and wisdom of having diverse, in the various senses of the term, teams in the workplace is being increasingly recognized, including in science and engineering (OSTP 2016). Furthermore, it would be prudent to establish systems for monitoring robot behavior. For example, the processes and outcomes of a medical robot's care could be evaluated using patient race or gender as the independent variable to ensure there are no significant differences found in the dependent variables of decisions made or quality of care when changing the value of this parameter.

Within the context of politics and other realms, a "litmus test" is a strategy for determining whether a candidate for a position (such as high office) is suitable for appointment or nomination. Oftentimes, it involves having a candidate answer a particularly contentious political question. Analogously, there may be ways to bring the litmus test strategy into the robotics realm. Researchers are starting to devise tests to assess whether machine learning algorithms are introducing gender or racial biases into the decision-making process by using word-associations (Bolukbasi et al. 2016). Other researchers are devising litmus-type tests for discrimination simply by analyzing the data going into a program and the decisions coming out the other end (Hardt et al. 2016). In this way, an application in the banking industry, for example, can be evaluated based on whether discriminatory lending practices are taking place (Devlin 2016). Transitioning to robotics, such a litmus-type test could thus be used to evaluate the existence of bias, for example, in a robot peacekeeper; one could examine whether the same decision would be made by the robot if its demographic input parameter with respect to race was modified.

What may complement the aforementioned efforts is to remove bias to the greatest extent possible from the words a robot selects while speaking with humans. For example, Caliskan et al. (2017) created the Word-Embedding Association Test (WEAT) and "applied it to a widely used semantic representation of words in AI, termed word embeddings." Their approach may provide insight into the bias hidden within word choices and perhaps eventually influence how to program robots.

An additional method for mitigating bias is to ensure that robots can articulate their reasoning process to users and others when making decisions. As the AI community develops explainable AI, an emerging field that seeks to enable machines to articulate their decision-making process (Castellanos 2016), these efforts could encompass methods applicable to robotics. By prompting robots to document the specific reasoning underlying a decision, others are then afforded a means to review the robot's logic with a critical eye toward identifying bias, or other flaws, before the robot commits itself to a decision. Lastly, case-based reasoning approaches could be used in which concrete examples showing positive and negative outcomes of bias are collected (Kuipers 2016). When a robot is required to make a new decision, the robot could compare its inputs and outputs to the different cases. Along with users, the robot could identify how closely its proposed course of action matches up to a moral exemplar.

Conclusion

The overarching goal of this paper is to bring the issue of bias more squarely to the attention of roboticists and others who are involved in designing robots. The robotics community needs to more systematically discuss the nature of bias and the role it has in influencing the design of next generation robots. Beyond describing how bias weaves its way into computing technologies, we sought to present strategies that could be utilized to mitigate or prevent bias in future robots. Admittedly, bias is difficult, if not impossible, to fully eradicate; yet the aforementioned strategies will hopefully give roboticists and others the tools to combat bias. Instead of reaffirming and more strongly entrenching bias, they could create robots that have beneficial, rather than detrimental, impacts on all segments of society.

References

American Association of Medical Colleges (AAMC). (2016). 2016 Physicians specialty data report. https://www.aamc.org/data/workforce/reports/457712/2016-specialty-databook.html. Accessed August 30, 2017.

American Civil Liberties Union. (2014). Racial disparities in sentencing. Submitted to the Inter-American Commission on Human Rights 153rd Session, October 27, 2014. https://www.aclu.org/sites/default/files/assets/141027_iachr_racial_disparities_aclu_submission_0.pdf. Accessed August 31, 2017.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed June 28, 2017.

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science, 355*(6323), 389–391.

Bogost, I. (2017). 'Artificial Intelligence' has become meaningless. *The Atlantic*, March 4. https://www.theatlantic.com/technology/archive/2017/03/what-is-artificial-intelligence/518547. Accessed August 30, 2017.

Bolukbasi, T., Chang K-W, Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. https://arxiv.org/pdf/1607.06520.pdf. Accessed July 3, 2017.

Borenstein, J., Herkert, J., & Miller, K. (2017a). Self-driving cars: Ethical responsibilities of design engineers. *IEEE Technology and Society Magazine, 36*(2), 67–75.

Borenstein, J., Howard, A., & Wagner, A. (2017b). Pediatric robotics and ethics: The robot is ready to see you now but should it be trusted? In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot ethics 2.0*. Oxford: Oxford University Press.

Brownstein, M. (2016). Implicit bias. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/win2016/entries/implicit-bias/. Accessed July 3, 2017.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.

Carpenter, J. (2015). Google's algorithm shows prestigious job ads to men, but not to women. *Independent*, July 7. http://www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-10372166.html. Accessed July 3, 2017.

Carty, S. S. (2011). Many cars tone deaf to women's voices. *Autoblog*, May 31. http://www.autoblog.com/2011/05/31/women-voice-command-systems/. Accessed February 4, 2017.

Castellanos, S. (2016). Capital One pursues 'explainable AI' to guard against bias in models. *The Wall Street Journal*, December 2. http://blogs.wsj.com/cio/2016/12/06/capital-one-pursues-explainable-ai-to-guard-against-bias-in-models/. Accessed February 10, 2017.

Chavalarias, D., & Ioannidis, J. P. A. (2010). Science mapping analysis characterizes 235 biases in biomedical research. *Journal of Clinical Epidemiology, 63*(11), 1205–1215.

Chawla, N. V., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16,* 321–357.

Chayes, J. (2017). How machine learning advances will improve the fairness of algorithms. Huffington Post, August 23. Accessed August 25, 2017.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies, 2015*(1), 92–112.

Devlin, H. (2016). Discrimination by algorithm: Scientists devise test to detect AI bias. *The Guardian*, December 19. https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-devise-test-to-detect-ai-bias Accessed February 9, 2017.

Executive Office of the President (EOP). (2016). Artificial intelligence, automation, and the economy. https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF. Accessed February 7, 2017.

Federal Trade Commission. (2016). Big data: A tool for inclusion or exclusion? Understanding the issues. *FTC report*. https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf. Accessed July 3, 2017.

Finley, K. (2016). Tech giants team up to keep AI from getting out of hand. Wired, September 28. https://www.wired.com/2016/09/google-facebook-microsoft-tackle-ethics-ai/. Accessed February 7, 2017.

Freeman, D. (2016). Self-driving cars could save millions of lives: But there's a catch. HuffPost Tech, February 18. http://www.huffingtonpost.com/entry/the-moral-imperative-thats-driving-the-robot-revolution_us_56c22168e4b0c3c550521f64. Accessed January 30, 2017.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies, 156,* 33–63.

Green, A., Carney, D., Pallin, D., Ngo, L., Raymond, K., Lezzoni, L., et al. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine, 22,* 1231–1238.

Griffiths, J. (2016). New Zealand passport robot thinks this Asian man's eyes are closed. *CNN*, December 9. http://www.cnn.com/2016/12/07/asia/new-zealand-passport-robot-asian-trnd/. Accessed February 4, 2017.

Guarino, B. (2016). Google faulted for racial bias in image search results for black teenagers. The Washington Post, June 10.

Hackett, R. (2016). Singapore is getting driverless taxi cabs. *Fortune*, April 5. http://fortune.com/2016/04/05/singapore-driverless-car-taxi-nutonomy/. Accessed February 8, 2017.

Hall, W., Chapman, M., Lee, K. M., Merino, Y. M., Thomas, T. W., Payne, B. K., et al. (2015). Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: A systematic review. *American Journal of Public Health, 105*(12), e60–e76.

Hardt, M. (2014). How big data is unfair: Understanding unintended sources of unfairness in data driven decision making, Medium, September 26. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de#.v96yl9fy6. Accessed January 28, 2017.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. October 11. https://arxiv.org/pdf/1610.02413.pdf. Accessed February 8, 2017.

Harpur, P. (2013). From universal exclusion to universal equality: Regulating ableism in a digital age. *Northern Kentucky Law Review, 40*(3), 529–565.

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43,* 274–306. doi:10.1111/j.1467-9833.2012.01565.x.

IEEE. (2017). Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. http://standards.ieee.org/develop/indconn/ec/ead_brochure.pdf. Accessed February 7, 2017.

Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology and Evolution, 28*(8), 474–481.

Kane, L. (2010). Exclusive ethics survey results: Doctors struggle with tougher-than-ever dilemmas. *Medscape*, November 11. http://www.medscape.com/viewarticle/731485. Accessed February 8, 2017.

Kang, C. (2016). The 15-Point federal checklist for self-driving cars. *The New York Times*, September 21.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI '15)* (pp. 3819–3828). New York, NY: ACM. doi:10.1145/2702123.2702520.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30,* 25–36.

Kuipers, B. (2016). Why and how should robots behave ethically? RoboPhilosophy 2016/TRANSOR 2016, Aarhus, Denmark. https://web.eecs.umich.edu/~kuipers/papers/Kuipers-robophilosophy-16-abstract.pdf. Accessed February 10, 2017.

Larson, S. (2016). Research shows gender bias in Google's voice recognition. *The Daily Dot*, July 15. http://www.dailydot.com/debug/google-voice-recognition-gender-bias/. Accessed February 4, 2017.

Levin, S. (2016). A beauty contest was judged by AI and the robots didn't like dark skin. *The Guardian*, September 8. https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people. Accessed February 4, 2017.

McFarland, M. (2016). Robot's role in killing Dallas shooter is a first. *CNN.com*, July 11. http://money.cnn.com/2016/07/08/technology/dallas-robot-death/. Accessed February 8, 2017.

McHugh, C., & Way, J. (2016). What is good reasoning? *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12299.

MIT Media Lab. (2017). MIT Media Lab to participate in $27 million initiative on AI ethics and governance. *MIT News*, January 10. http://news.mit.edu/2017/mit-media-lab-to-participate-in-ai-ethics-and-governance-initiative-0110. Accessed February 7, 2017.

Muench, U., Sindelar, J., Busch, S. H., & Buerhaus, P. I. (2015). Salary differences between male and female registered nurses in the united states. *JAMA, 313*(12), 1265–1267. doi:10.1001/jama.2015.1487.

Nwana, H. S. (1996). Software agents: An overview. *Knowledge Engineering Review, 11*(3), 1–40.

Office of Science and Technology Policy (OSTP). (2016). Reducing the impact of bias in the stem workforce: Strengthening excellence and innovation. A Report of The Interagency Policy Group on Increasing Diversity in the Stem Workforce by Reducing the Impact of Bias.

Otterbacher, J. (2016). New evidence shows search engines reinforce social stereotypes. Harvard Business Review, October 20. https://hbr.org/2016/10/new-evidence-shows-search-engines-reinforce-social-stereotypes. Accessed February 4, 2017.

Pulliam-Moore, C. (2015). Google photos identified black people as 'gorillas,' but racist software isn't new. *Fusion*, July 1. http://fusion.net/story/159736/google-photos-identified-black-people-as-gorillas-but-racist-software-isnt-new/. Accessed February 4, 2017.

Robinson, D., & Koepke, L. (2016). Stuck in a pattern: Early evidence on "predictive policing" and civil rights. A report from Upturn. https://www.teamupturn.com/reports/2016/stuck-in-a-pattern. Accessed January 28, 2017.

Rodger, J. A., & Pendharkar, P. C. (2004). A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies, 60*(5–6), 529–544.

Rose, A. (2010). Are face-detection cameras racist? *Time*, January 22. http://content.time.com/time/business/article/0,8599,1954643,00.html. Accessed February 4, 2017.

Šabanović, S., Chang, W-L, Bennett, C. C., Piatt, J. A., & Hakken, D. (2015). A Robot of my own: Participatory design of socially assistive robots for independently living older adults diagnosed with depression. In: *Human aspects of IT for the aged population: Design for aging. Lecture Notes in Computer Science* (Vol. 9193, pp. 104–114).

SAE International. (2016). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles J3016. http://standards.sae.org/j3016_201609/. Accessed January 17, 2017.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28.

Steele, C. M. (2010). *Whistling Vivaldi: And other clues to how stereotypes affect us (issues of our time)*. New York, NY: W. W. Norton and Co.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., et al. (2016). Artificial intelligence and life in 2030. One hundred year study on artificial intelligence: Report of the 2015–2016 study panel. Stanford, CA: Stanford University. http://ai100.stanford.edu/2016-report. Accessed February 7, 2017.

Tatman, R. (2016). Google's speech recognition has a gender bias. Making Noise and Hearing Things, July 12. https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/. Accessed February 9, 2017.

Toor, A. (2016). Drones will begin delivering blood and medicine in the US. *The Verge*, August 2. http://www.theverge.com/2016/8/2/12350274/zipline-drone-delivery-us-launch-blood-medicine. Accessed February 8, 2017.

Wagner, L. (2015). North Dakota legalizes armed police drones. *The Two-Way*, August 27. http://www.npr.org/sections/thetwo-way/2015/08/27/435301160/north-dakota-legalizes-armed-police-drones. Accessed February 8, 2017.