Unintended Consequences of Biased Robotic and Artificial Intelligence Systems

By Ludovic Righetti, Raj Madhavan, and Raja Chatila

The IEEE Robotics and Automation Research and Practice Ethics Committee (RARPEC) is intended as a platform to exchange ideas and discuss the impacts and practice of robotics and automation (R&A) technologies in research, development, and deployment that appear to pose ethical questions for humanity. With increased awareness and controversies surrounding R&A, RARPEC is publishing a series of opinion pieces that will focus on separating hype from reality by providing an objective and balanced treatment of technological, ethical, legal, and societal perspectives. Third in the series, this piece focuses on the implications and consequences of bias in robotic systems. Please send your feedback and suggestions to the chair of the committee, Raj Madhavan, at raj.madhavan@ieee.org. We look forward to your comments!

It seems natural to assume that well-engineered robots and artificial intelligence (AI)-based systems will always behave in an unbiased and nondiscriminatory manner when they interact with or provide services to people. As long as engineers are not malicious, won't robots perform tasks equally well, independent of one's gender, ethnicity, age, or socioeconomic background?

While this could be considered a common-sense assumption, a growing body of work shows that algorithms often exhibit unintended and unexpected biases, leading to unfair discrimination against certain groups of people. Although seldom addressed in robotics, these issues will certainly permeate all applications that involve interactions with humans and have recently raised concerns in the machine-learning community. In this article, we discuss some of these issues, including how they could impact robotics and automation and potential approaches to ensure the fair deployment of robotics technologies.

Digital Object Identifier 10.1109/MRA.2019.2926996
Date of publication: 5 September 2019

When Algorithms Reinforce or Lead to Discrimination

Commercial computer-vision systems are prime examples of systems that unfairly discriminate or contribute to reinforcing stereotypes due to misclassification errors. In 2015, it was reported that Google's photo service classified black people as gorillas in certain pictures, reinforcing a racist stereotype. Surprisingly, the problem appeared difficult to resolve: more than two years later, the seemingly preferred solution was to simply remove the gorilla category from the classification algorithm [1]. An audit of several commercial facial-recognition systems for gender classification revealed that every tested system performed systematically worse with women and, in particular, with dark-skinned women as compared with light-skinned men [2]. Amazon's Rekognition system had 30% more classification errors for dark-skinned women while close to perfect classification performance for light-skinned men. Interestingly, the audit, which publicly named some of the tested companies, had a positive impact. Several months later, those companies had released new versions of the algorithms that demonstrated a significant decrease in performance disparities. These examples shed light on how seemingly value-free technologies can lead to racial or gender-based inequities, as technology-based services may not function equally well for everyone. This finding raises serious concerns because facial recognition is being widely deployed, particularly for government services and law-enforcement applications.

However, the problem goes beyond mere computer vision. For example, Amazon reportedly developed a hiring tool to help rank candidates using data from previous hires [10]. The system was shown to systematically downgrade candidates who attended all-women colleges (and, generally, all resumes containing the word *women*). Despite efforts to remove such bias, it seems that the hiring tool was finally abandoned. Interestingly, gender-based discrimination appears to be difficult to prevent due, among other causes, to a history of gender-biased hiring practices that permeate the data. This illustrates how an algorithm can potentially introduce or reinforce, and indefinitely perpetuate, already rampant discriminatory practices.

A more disconcerting example is the infamous COMPAS algorithm that certain courts in the United States used to assess the likelihood that a criminal defendant becomes a recidivist.

A ProPublica study found that "black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk" [3]. These tools go beyond mere image classification and influence decisions about sentencing and parole, with important impacts on people's lives and freedoms. In addition to reinforcing discriminatory practices, the company that created the algorithm was criticized for lacking transparency because the algorithm is unknown and does not allow for external auditing to understand and potentially correct biased outcomes. Search engines have also been accused of reinforcing bias and stereotypes. For example, a study of the Google Ads system found that "setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male" [4].

Beyond data sets, the very nature of classification algorithms, especially those based on visual data, has raised ethical concerns [5]. The mere decision to define classes can already be discriminatory. For example, when using visual data to find someone's gender, algorithms typically employ binary classification based on biological sexual traits (male or female) expressed as visual features. This selection of binary classes implicitly denies more complex gender identities with the risk of invisibilizing or discriminating against gender-fluid or transgender people. Furthermore, it has been argued that reducing personality traits, gender, or emotions to visual features is dangerous, as it potentially "makes a value judgement about a person based on their mannerisms and the appearance of their body" [5]. Recent applications of machine learning have raised important concerns about technology enforcing prejudicial practices when aiming to recognize criminals, sexual preferences, or complex emotions based solely on visual data. These examples shed light on the potentially complex socioeconomic and cultural side effects of seemingly value-free applications of recent AI-based technologies.

What Are the (Unintended) Consequences for Robotics?

Many of these technologies are, in fact, core building blocks of robotic systems intended to be used in interactions with humans. It therefore is clear that unsuspected and unintended bias issues will also be pervasive in robotics applications. A perception system for an autonomous car that detects people in a significantly different way based on their skin color or gender would create an unfair distribution of risks across the population, with certain groups of people more likely to enter into an accident than others. Light-skinned versus dark-skinned issues have already come to the fore when autonomous vehicles try to classify what constitutes a human on the road [11]. However, the particular complexity of robotic systems might also accentuate these issues because multiple biased algorithms could reinforce one another, severely increasing the difficulty of detecting bias during complete system operation. Howard and Borenstein [6] have discussed in depth how bias in robotics could lead to the unfair treatment of certain demographic groups. They question how life-critical decisions will be made not only by self-driving cars but also by medical or service robots if algorithms introduce or reinforce certain discriminatory practices. Furthermore, they argue that algorithms could deepen unfair surveillance and profiling of minorities by a robotization of law enforcement systems, e.g., using computer vision to assist policing practices could accentuate unfair targeting.

Applications involving human–robot interactions are especially vulnerable, as the feedback nature of such interactions could amplify biased responses of the robotic device in ways that are difficult to detect. In particular, if the interaction creates an issue, it might be difficult to distinguish detrimental behaviors before field deployment. While trust is an important and desirable factor for human–machine interactions, strong user confidence and reliance on robotic systems could potentially worsen the detection of unfair behaviors. One could legitimately ask whether biased machines will not only perpetuate the discriminatory practices present in all societies but also introduce unforeseen problems that will make them even worse—and at an unprecedented scale—as robotic technologies increasingly interact with humans.

The Need for Multidisciplinary Approaches to the Problem

The reasons for biased outcomes often are very complicated. A lack of diversity in the training data sets used in machine learning is a central cause. For example, the Labeled Faces in the Wild data set contains more than 15,000 images of faces but as few as 7% of the images are of black people [7]. It has been argued that this lack of data diversity stems, to some extent, from the actual lack of diversity in many AI-related fields, creating a feedback loop of bias [5]. The available data also can be intrinsically skewed and not usable, as in the hiring tool example. It is not clear how certain data sets can be diversified, because historic data may contain the stereotypical patterns that one would expect to correct. Furthermore, explicitly removing the biasing attributes from the learning algorithm might not be sufficient due to the correlations with other features that are often hidden, highlighting the additional difficulty of choosing the appropriate input features. Moreover, we all have unconscious or implicit biases that might further influence algorithm design and outcomes.

The consensus in the machine-learning community is that bias is an issue that needs to be specifically addressed. From a technical standpoint, research on algorithmic fairness proposes formal methods to mitigate the risk of bias. For example, anticlassification strategies ensure that classification results do not depend on protected attributes, such as gender or race, and classification parity ensures that evaluation performance is equal across groups defined by protected attributes. However, there are fundamental technical limitations to a mere algorithmic approach [8]. Indeed, it has been

shown that not all proposed fairness criteria can be optimized at the same time, therefore limiting guarantees that an algorithm will always be fair. An adequate enumeration of protected attributes and their proxies (e.g., ZIP codes are correlated with race in the United States, thus leading to redlining practices that exacerbate inequalities) might be difficult without, for example, working with sociologists who can help put real-world abstract data into categories and contextualize data in particular socioeconomic and cultural environments. Several companies are releasing fairness tool kits to be integrated into software-development pipelines early in the process. Industrial standards, such as the IEEE Ethics Certification Program for Autonomous and Intelligent Systems within the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, are also being proposed [12]. However, to date, the methodology for certification is not known, and it is not clear whether such certification is enough to truly avoid biased systems.

More holistic approaches that encompass a wider range of disciplines have recently been advocated as necessary during the design of complex AI or robotic systems [9] as a way not only to provide technical solutions but also to better understand the socioeconomic and cultural contexts of use of the technology. This includes taking special care in the construction, documentation, and validation of data sets and making the push for more transparent algorithms that can be monitored and audited during real-world operations, perhaps by independent institutions.

Wide societal acceptance and trust of robotic systems will require a concerted effort involving sociologists, ethicists, philosophers, and technologists to ensure fairness and build trust in deploying autonomous and interactive systems. Perhaps Mark Twain wasn't thinking about bias in robotics, but he said it best when he noted, "What gets us into trouble is not what we don't know. It's what we know for sure that just ain't so"!

References

[1] J. Vincent, "Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech," *The Verge*, Jan. 12, 2018. [Online]. Available: https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

[2] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products," in *Proc. AAAI/ACM Conf. AI Ethics and Society*, 2019. [Online]. Available: http://www.aies-conference.com/wp-content/uploads/2019/01/AIES-19_paper_223.pdf

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, May 23, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A Tale of opacity, choice, and discrimination," in *Proc. Privacy Enhancing Technologies*, 2015, pp. 92–112. doi: 10.1515/popets-2015-0007.

[5] S. Myers West, M. Whittake, and K. Crawford, "Discriminating systems: Gender, race and power in AI," AI Now Institute, New York Univ., New York, 2019. [Online]. Available: https://ainowinstitute.org/discriminatingsystems.pdf

[6] A. Howard and J. Borenstein, "The ugly truth about ourselves and our robot creations: The problem of bias and social inequity," *Sci. Eng. Ethics*, vol. 24, no. 5, pp. 1521–1536, Oct. 2017. doi: 10.1007/s11948-017-9975-2.

[7] H. Han and A. K. Jain, "Age, gender and race estimation from unconstrained face images," Michigan State Univ., East Lansing, Tech. Rep. MSU-CSE-14-5, 2014. [Online]. Available: http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

[8] J. Kleinberg, S. Mullainathan, and M. Raghavan, Inherent trade-offs in the fair determination of risk scores. 2016. [Online]. Available: https://arxiv.org/abs/1609.05807

[9] M. Whittaker et al., "AI now report," AI Now Institute, New York Univ., New York, 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf

[10] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, Oct. 9, 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[11] B. Wilson, J. Hoffman, and J. Morgenstern, Predictive inequity in object detection. 2019. [Online]. Available: https://arxiv.org/abs/1902.11097

[12] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethics in action." [Online]. Available: https://ethicsinaction.ieee.org/

FROM THE EDITOR'S DESK *(continued from page 4)*

of the reasoning process was an essential part of AI expert systems in the 1980s, the research domain of explainable AI is still in its infancy for systems that learn from so-called big data. A thorough knowledge of AI algorithms is necessary to guarantee aspects of justice toward individuals and groups as well as to prevent bias.

As usual, this September issue of *IEEE Robotics and Automation Magazine* includes articles submitted on a variety of topics rather than focusing on a particular theme, as do special issues. Calls for upcoming special issue papers focus on deep learning and soft robotics. Please check the Society website for more information. To support the reproducibility of robotics and automation research, the IEEE Robotics and Automation Society (RAS) waives, for two years and a maximum of five articles per year, open access fees for reproducible articles (designated *R-Articles*), the first of which appears in this issue. Enjoy your reading!