Bias In, Bias Out - Diversity In, Diversity Out

Karime Pereida and Melissa Greeff

Abstract— **Roboticists develop technologies that are used by people worldwide, consequently impacting many aspects of human life - from healthcare and law enforcement to autonomous transportation. The development of these technologies involves design and innovation - both of which rely on personal choice and experience. Hence, personal biases, whether intentionally or unintentionally, tend to be embedded in the final product designs. Homogeneous teams of designers and engineers are more likely to develop products that overlook the needs of a given part of the population - even missing gaps for potential technological innovation. In this talk we emphasize some of the negative impacts a lack of diversity has on robotic innovation by highlighting examples of embedded biases within certain technologies and providing some evidence that this is linked to a lack of diverse teams. If our aim as a community is to increase research capacity, creativity, and broaden the impact of robotics, making it a more diverse field must be a goal.**

I. Aspirational Robotics?

Robotics is a growing engineering field that promises to revolutionize the way we live. It has the potential to impact every aspect of our lives ranging from entertainment to health care and policing. We are faced with a choice. Do we build and design robots that are representative of our world as it currently is? Or, do we build robots for a world that many aspire to? While for many of us technology-enthusiasts the latter may seem like an obvious choice, this choice means that we need to design robots and/or algorithms that are able to overcome long-standing societal injustices. We do not have 'aspirational' data to design these algorithms and robots from. We will argue that instead we may need to rely on 'aspirational' design - done by innovators from a diversity of backgrounds and experiences. We do not argue diversity for fairness (a term with many definitions [1]). Rather, we argue that it is a potential solution to both prevent our technology from reinforcing current societal biases and to increase research capacity and the impact of robotics [2].

II. To Bias or Not To Bias

Biases are embedded within the make-up of our society and consequently also within our current data. Bias can refer to either the behaviors that can contaminate research projects or to performance problems directly correlated with class-imbalance in datasets used to prove a hypothesis or to train algorithms [3]. As humans, we may not be consciously aware of our biases, which is a phenomenon referred to as *implicit bias* [4]. However, intentionally or unintentionally, as roboticists, we inflict our biases onto robots. This affects both their hardware and software design.

The authors are with the Dynamic Systems Lab (www.dynsyslab.org) at the University of Toronto Institute for Aerospace Studies (UTIAS), Canada. Email: karime.pereida, melissa.greeff@robotics.utias.utoronto.ca

Biases are present in the hardware of the robots in two main ways: (i) how humans perceive robots based on their appearance, and (ii) which physical characteristics have been given to robots that carry out certain tasks. Most of the home robots are designed in white plastic or a metallic finish. A study by professor Bartneck at the Human Interface Technology Lab at the University of Canterbury showed that participants have a bias against robots with black plastic and were more likely to shoot black robots that posed no threat versus white robots [5]. Biases extend to the perceived gender of robots. In Singapore, a study of 198 young adults showed that participants felt more comfortable with the idea of a male security robot and a female housecleaning robot [6].

Currently, artificial intelligence (AI, used as an umbrella term) is the main mechanism that powers the decision-making process of an agent regardless of whether this agent has a robot body or not. Bias is often encoded in learning algorithms and manifests itself in different ways, mostly in ways that perpetuate these biases. As these algorithms are released into the world, the impact of bias increases. Studies of biases of AI based on race and gender have surfaced over the last few years. Search engines were noted to deliver job postings for well-paying technical jobs to men but not to women [7]. Searching the keywords 'black teenagers' provided images of mugshots of black teenagers while searching for 'white teenagers' provided images of happy white teenagers [8]. In general, face recognition performs better with males of lighter skin tones and has problems recognizing black females [9]. In particular, Google Photos labelled black people as gorillas; the response of the company was to ban terms such as 'gorilla' and 'chimpanzee' while they worked on longer term fixes [10]. Finally, a female of Taiwanese decent complained that her camera kept labelling her as 'blinking' [3].

Speech recognition algorithms have also been studied and instances of bias have been detected over the years. Medical voice-dictation software could more accurately recognize input from a man versus a woman [11]. 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 [3]. Moreover, the gender bias has been widespread as artificial intelligence assistants like Siri, Alexa and Google Assistant have been given female voices.

In recent years there has been a push toward de-biasing these types of algorithms. We name four interesting ideas - the first two target algorithm structure, the last two target data structure. The first idea is to adjust the learning algorithm

to reduce its dependence on attributes that affect it from achieving equitable performance across subpopulations [12]. A similar idea tries to achieve equitable performance by maximizing accuracy under a 'fairness' constraint [13]. A third idea is to use machine learning itself to identify and quantify bias in data [14]. The fourth idea used by the Data Nutrition Project [15] instead targets 'unhealthy' data by providing a diagnostic tool to assess data.

While a step in the right direction, these technical approaches still suffer from two main limitations. Firstly, they require potential sensitive data, for example ethnicity, to be accurately recorded. Secondly, they still require designers to decide a priori what types of biases they want to to avoid and to what extent. Consequently, the implicit bias of the designer is still embedded in the de-biasing mechanism!

The effects of the above biases range from minor inconveniences as not being able to turn up the volume of the radio in a car to having major life impacts such as influencing the quality of education or healthcare that an individual or group receives [3]. In recent years, the influence of AI has reached areas such as hiring, housing, criminal justice and the military. This technology has automated biases of designers such as devaluing women's resumes, perpetuating employment and housing discrimination, and enshrining racist policing practices and prison convictions [16]. Professor Bartneck argues that if the field of robotics does not incorporate diversity now, it will suffer the same issues that established industries are currently trying to correct [3].

III. Organizations and Diversity

People who are most negatively affected by the above mentioned biases are women and people of color. This correlates highly with representation of these groups in various engineering venues. Women account for only 18% of authors at leading AI conferences, 20% of AI professorships, and 15% and 10% of research staff at Facebook and Google, respectively. Racial diversity is even worse: black workers represent only 2.5% of Googles entire workforce and 4% of Facebooks and Microsofts. There is no data available for transgender people and other gender minorities [16].

Many organizations have started to foster diversity in robotics and related fields. For example, Code2040 is a non-profit that seeks to foster racial equity in tech by enhancing participation and leadership of black and Latin technologists in the innovation economy [17]. The Algorithmic Justice League aims to highlight algorithmic bias, provide a space for people to voice concerns, and develop practices for accountability [18]. Conferences such as Robotics, Science and Systems have launched programs such as 'Inclusion@RSS' to increase the participation of groups traditionally underrepresented in robotics. However, more work needs to be done. Overall, inclusion methods that increase the number of candidates from underrepresented groups tend to underestimate other systemic disadvantages that prevent women and minorities from staying in the field, such as harassment, unfair compensation, and imbalances of power [16].

IV. Conclusion

In this paper we have discussed the varied effect of bias on many of our current robotic solutions. We've proposed that 'fixing' the data is not enough - instead a more robust solution requires greater diversity within the team of designers. This is the bottleneck if we want to design robots that do not perpetuate long-standing societal bias. While there are many organizations with varying diversity missions, this still remains an open challenge. We have motivated the need to make this a community goal - we now look for more creative ways to address this aim.

References

[1] A. Narayanan. Tutorial: 21 fairness definitions and their politics. Youtube. [Online]. Available: https://www.youtube.com/watch?v=jIXIuYdnyyk

[2] W. S. Journal. The need for diversity in robotics. Youtube. [Online]. Available: https://www.youtube.com/watch?v=AVUzgSe0b3w

[3] A. Howard and J. Borenstein, "The ugly truth about ourselves and our robot creations: the problem of bias and social inequity," *Science and engineering ethics*, vol. 24, no. 5, pp. 1521–1536, 2018.

[4] J. Borenstein, J. Herkert, and K. Miller, "Self-driving cars: Ethical responsibilities of design engineers," *IEEE Technology and Society Magazine*, vol. 36, no. 2, pp. 67–75, 2017.

[5] C. Bartneck, K. Yogeeswaran, Q. M. Ser, G. Woodward, R. Sparrow, S. Wang, and F. Eyssel, "Robots and racism," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 196–204.

[6] T. Love. (2018) Robots have a diversity problem. [Online]. Available: https://medium.com/s/thenewnew/robots-the-new-identity-politics-4b36700630db

[7] J. Carpenter. (2015) Google's algorithm shows prestigious job ads to men, but not to women. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-10372166.html

[8] B. Guarino. (2016) Google faulted for racial bias in image search results for black teenagers. [Online]. Available: https://www.washingtonpost.com/news/morning-mix/wp/2016/06/10/google-faulted-for-racial-bias-in-image-search-results-for-black-teenagers/?utm_term=.2df65baa226f

[9] T. Simonite. (2018) Photo algorithms id white men fineblack women, not so much. [Online]. Available: https://www.wired.com/story/photo-algorithms-id-white-men-fineblack-women-not-so-much/

[10] ——. (2018) When it comes to gorillas, google photos remains blind. [Online]. Available: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[11] J. A. Rodger and P. C. Pendharkar, "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*, vol. 60, no. 5-6, pp. 529–544, 2004.

[12] C. Dwork, M. , Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 2012, pp. 214–226.

[13] M. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness constraints: A mechanism for fair classification," 2015. [Online]. Available: https://arxiv.org/pdf/1507.05259.pdf

[14] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, pp. E3635–E3644, 2018. [Online]. Available: https://www.pnas.org/content/115/16/E3635

[15] The data nutrition project. MIT Media Lab. [Online]. Available: https://datanutrition.media.mit.edu/

[16] K. Hao. (2019) Ais white guy problem isnt going away. [Online]. Available: https://www.technologyreview.com/s/613320/ais-white-guy-problem-isnt-going-away/

[17] Code2040. [Online]. Available: http://www.code2040.org/

[18] Algorithmic justice league. [Online]. Available: https://www.ajlunited.org/