

# The impact of gender and race bias in AI

August 28, 2018, Analysis / Autonomous Weapons / Gender / Law and Conflict / Weapons

Noel Sharkey



Automated decision algorithms are currently propagating gender and race discrimination throughout our global community. The major culprits are machine learning with big data. Drawing on the lessons learned, the question asked here is, can we take away meaningful human control and put our trust in artificial intelligence (AI) to select targets and attack them with violent force?

To answer this, we take a journey into less deadly application domains where the biases of such systems are clear to see. En route we'll examine the causes of bias inherent in the use of machine learning with big data and why a lack of transparency makes the bias problem difficult to fix. Finally, similar problem

of bias in the use of automated face recognition are discussed in relation to target acquisition.

Much of the recent discussions about autonomous weapons systems (AWS) have focused on careful military planning, on maintaining wider loops for human control and on human oversight. While conventional military planning is important, we must consider the stark reality of other ways that the technology may be used. Modern armed conflict is rarely about high densities of military personnel meeting on a battlefield with the aim of killing as many of their opponents as they can. It is replete with un-uniformed combatants moving through and hiding among civilian populations. **Some believe that AWS will be able to pick out specific people or classes of people who may be legitimate targets of attack (setting aside the plethora of international human rights issues). But is this a humanitarian disaster waiting in the wings?**

## Bias in decision algorithms

The acceleration in the commercialisation of AI has largely been the product of two main factors (i) an increase in power, memory and speed of computers and (ii) the availability of large quantities of data about so many aspects of our lives. These factors have enabled the use of machine learning tools in an unprecedented way for innovative commercial applications. These include algorithms being delegated with decisions that impact on people's lives in the belief that algorithms will make more objective and hence fairer and less biased decisions than error prone and biased humans

Unfortunately, it is becoming alarmingly clear from a growing body of evidence that decision algorithms are perpetuating injustice for many. Many cases are emerging that postcode, ethnicity, gender, associations and poverty negatively bias the decisions being delegated to machines. Tech giant such as *Google and Microsoft* openly admit that there are bias problems and they have been doing their best, unsuccessfully so far, to find solutions. Areas of injustice coming to light include: Insurance risk, mortgage loan decisions, shortlisting, hiring and interviewing for jobs, job advertising, bail and recidivism risk, assessment, custodial sentencing, airport security and predictive policing.

One of the common causes of decision biases arises from ethnicity or gender biases, often unconscious, of the programmer or of those classifying the data samples. Another major cause is the application of machine learning algorithms, such as deep learning, and the way in which they are trained. I will cover two of the inherent causes below and then explain why lack of transparency makes them difficult to fix

## How machine learning can ignore minorities

One of the benefits of using machine learning systems in an engineering context is that they reduce or remove the impact of outliers (examples outside of the norms in the data) in the training data. For example, shaky arm movements of a robot can be turned into smooth movements by training with machine learning. However, in the context of a decision algorithm, the 'outliers' can be the minority groups in the data. These may be people from different ethnic groups or a low represented gender. *Google illustrates* this problem by training a system with drawings of shoes. Most of the sample are the

plain shoes or trainers drawn by the majority of people. But some people drew high-heels. Because this was a minority, the post training tests misclassified high heels as 'not shoes'. This is a simple example where the misclassification can be clearly seen. In a targeting context there would be no way of knowing which minorities could mistakenly fall into the category of legitimate targets. And in conflict it is rarely possible to get clear information about casualties and whether or not they were legitimate targets or collateral damage.

## **Bias problems in the data**

Our societal values and norms are constantly evolving with massive changes over the last 50 years in what is acceptable to say in the media or in the workplace and also as to what are discriminatory practices. However, it appears that most of the old values are locked into the internet where much of the training data for machine learning algorithms are derived. Studies have demonstrated that, for example, 'man' is associated with boss, president, leader and director, whereas 'woman' is associated with helper, assistant, employee and aide. Google searches for black names like Leroy and Keisha yield ads associated with crime, whereas white names like Brad and Katie yield sites with contact details. Searches for professional hair yield images of white women whereas for unprofessional hair we get images of well-groomed black women. Then there are latent biases in data created by frequency of occurrence—e.g., searching on Google images for pictures of great physicists yields mostly pictures of men. The societal push towards greater fairness and justice is being held back by historical values about poverty, gender and ethnicity that are ossified in big data. There is no reason to believe that bias in targeting data would be any different or any easier to find.

## **Lack of transparency and explanation**

In the Google shoe classification example described above—amplification through filtering—the data may contain flat soles, small heels, high heels, laces, no laces, strips, shiny material etc. When the algorithm has completed learning, it would be difficult or perhaps impossible to work out which features of the drawing had been selected to trigger the shoe/non-shoe decision. This lack of transparency is because the end result of learning is large matrices of numbers that are used to generate the decisions. No one has, as yet, found a general means to probe the matrices to find out what features of the data it has assigned to the decision. This means that no explanation can be given for why a particular decision has been made. Imagine the dangers that this poses for targeting decisions where we cannot tell what features are responsible for classifying a person or object as a legitimate target. This is the dark recess where bias lives.

## **The consequences of bias in armed conflict**

All of this injustice is happening in the civilian domain where there has been plenty of time to test, scrutinize and construct these systems. It emphasises the problems of accountability and responsibility in the automation of weaponry that we have been discussing for years.

When it comes to armed conflict, data management will be considerably more difficult. If there is bias in civil society, just imagine the kinds of bias that will be built into algorithms being delegated with life and death decisions in conflict zones where we have little or no proper understanding of the cultures involved. Given the enormous challenges we are already facing in trying to develop fair and just AI systems, the chances of justice in civilian laden conflict zones is vanishingly small.

## There is even bias in automated face recognition

Proponents of autonomous weapons systems might still argue that automated face recognition could form an objective and unbiased basis for automating kill decisions in urban warfare. They may even suggest that it is the ideal tool to track down and kill 'high value targets'. After all, it has been shown to be accurate in laboratories and it has been developed and sold by the major tech companies like Amazon, Google and Microsoft. But there are problems.

First, the NGO, Big Brother Watch used freedom of information requests to obtain accuracy data on the UK police force's use of NEC's NeoFace Watch to find criminal faces among crowds. The results of their ensuing *report* were shocking. The average false face recognition was 95%. Yes, that means that only 5% of those identified were that person. The worst results were the metropolitan police force's use of the technology at the big Afro-Caribbean Notting Hill Carnival with only 2% correct recognition accuracy over a weekend. Just imagine the humanitarian consequences if that had been an AWS selecting targets

Second, the American Civil Liberties Union (ACLU) conducted a *range of tests* with Amazon's Rekognition system that is becoming popular among U.S. police departments. One of their tests matched photographs of members of the U.S. Congress against a database of 25,000 publicly available 'mug shots' of criminals. The system incorrectly matched 28 members of Congress with people who have been arrested. This has led U.S. lawmakers to raise questions about the police use of the technology and both *Amazon* and *Microsoft* (who make their own facial recognition software) have called for new regulations. From the perspective of bias, it is important to note that the ACLU test showed that a disproportionate number of African-American and Latino members of Congress were misidentified as criminals.

The best results reported for face recognition are for *white males* and that has been the conclusion of a number of academic studies. Gender, age and shade of skin really do matter to automated face recognition. Joy Buolamwini from MIT, tells the story of how her research on using a computer avatar was hampered because the face recognition software could not even find her face, never mind recognise it—the *missing face problem*. She had to wear a white mask to be seen.

Buolamwini then conducted a *comprehensive research study* on three major commercial face recognition systems for gender classification: IBM Watson Visual Recognition, Microsoft Face Detect and Face++. The face pictures were of parliamentarians from three northern European countries and three African countries. A range of skin shades were represented. The result was that all systems perform better on males versus females and the lighter the skin the more accurate the result. All performed worst on darker female faces with error rates from 20.8% to 34.7%.

## In conclusion

It should be clear from evidence presented above that both AI decision algorithms and face recognition algorithms can be alarmingly biased or inaccurate with darker shades of skin and with women. These may well improve over time but there have been no magic bullet solutions despite massive efforts and several announcements. Many of the companies developing software, particularly for policing, insist that they did well on their inhouse testing. It has remained for other organisations, such as NGOs, to collect the data and demonstrate the biases, yet the systems keep on getting rolled out. It is the familiar old story that once there has been huge investment in a technology it continues to be used despite its failings. Let us not make the same mistake with targeting technology.

Discriminatory systems are bad enough in the civilian world where new cases of injustice to women and people with darker shades of skin are turning up almost weekly. But while it can be difficult for those who suspect discrimination to take legal action, there is at least the potential to reverse such unjust decisions. It is a different story when dealing with the technologies of violence. Once someone has been misclassified and targeted with lethal force by an unfairly biased decision process, there is no overturning the decision.

Technology, and particularly AI, has always gotten ahead of itself with ambition outstripping achievement. In my long experience working on the subject and reviewing many research proposals, ambition often wins the day. Indeed, ambition is often a positive step towards achievement. In many cases it can still be worthwhile even if the achievement falls well short of the ambition. However, when it comes to technologies of violence, we need to be considerably more cautious of ambitious claims about speculative technology that can lead us down the wrong path.

Like a retired police horse, it is time to take off the blinkers and look at the current state of technology and its problematic relationship to the technologies of violence. We cannot simply ignore the types of discriminatory algorithmic biases appearing in the civilian world and pretend that we can just make them go away when it comes weapons development and use. These are just some of the problems that have come to light, since the increased use of AI in society. We don't know what further problems are around the corner or what further biases are likely to occur in targeting technologies.

The moral of this tale is simple. We must take a precautionary approach to the use of AI in weapons technology and AWS in particular. We must not rely on the possibility of future fixes but instead make decisions based on what the technology is capable of today. It is time now for nation States to step up to the mark and begin negotiations for a new international legally binding instrument to ensure the meaningful human control of weapons systems is preserved.

\*\*\*

## Related blog posts

- *The human nature of international humanitarian law* Eric Talbot Jensen

- *Autonomous weapons: Operationalizing meaningful human control* Merel Ekelhof
- *Human judgment and lethal decision-making in war* Paul Scharre
- *Autonomous weapon and human control* Tim McFarland
- *Autonomous weapon systems: An ethical basis for human control?* Neil Davison
- *Autonomous weapon systems: A threat to human dignity?* Ariadna Pop
- *Ethics as a source of law: The Martens clause and autonomous weapons* Rob Sparrow
- *Autonomous weapons mini-series: Distance, weapons technology and humanity in armed conflict* Alex Leveringhaus
- *Introduction to Mini-Series: Autonomous weapon systems and ethics*





\*\*\*

## Key ICRC documents on AWS

- *VOX News/IRC Video clip* on autonomous weapons, May 2018
- *ICRC Statement* to UN CCW Group of Governmental Experts (GGE), April 2018
- *ICRC Statement* to UN CCW Group of Governmental Experts (GGE), November 2017
- *ICRC Report* on Ethics and autonomous weapon systems, April 2018
- *Paper* on autonomous weapon systems under international humanitarian law, November 2017
- *ICRC Expert meeting report*, 2016
- *ICRC Expert meeting report*, 2014

*DISCLAIMER: Posts and discussion on the Humanitarian Law & Policy blog may not be interpreted as positioning the ICRC in any way, nor does the blog's content amount to formal policy or doctrine, unless specifically indicated.*

Tags: AI, algorithms, Amazon, artificial intelligence, big data, discrimination, face recognition, gender, Google, IHL, law of armed conflict, LOAC, machine learning, Microsoft, race

Share this article    

## Comments

ALEJANDRA VILLALOBOS RUIZ, 15 October 2018

This is terrifying to say the least, again the funding and economic value are placed before the reality of the effects on civilian communities.

*Leave a comment*

Name \*

Email address \* *This is for content moderation. Your email address will not be made public.*

Your comment

**Post comment**



*Get Law & Policy updates*

Adobe Acrobat

Conversion Cancelled

Convert webpage to PDF...

Add to existing PDF...

Visit Document Cloud

Open PDF in Acrobat

Preferences...