Google's new principles on AI need to be better at protecting human rights

June 18 2018, by Lorna Mcgregor And Vivian Ng

There are growing concerns about the potential risks of AI – and mounting criticism of technology giants. In the wake of what has been called an AI backlash or "techlash", states and businesses are waking up to the fact that the design and development of AI have to be ethical, benefit society and protect human rights.

In the last few months, Google has faced protests from its own staff against the company's AI work with the US military. The US Department of Defense contracted Google to develop AI for analysing drone footage in what is known as "Project Maven".

A Google spokesperson was reported to have said: "the backlash has been terrible for the company" and "it is incumbent on us to show leadership." She referred to "plans to unveil new ethical principles." These principles have now been released.

Google's chief executive, Sundar Pichar, acknowledged that "this area is dynamic and evolving" and said that Google would be willing "to adapt our approach as we learn over time." This is important because, while the principles are a start, more work and more concrete commitments are needed if Google is going to become effective in protecting human rights.

Google's principles on AI

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.
7. Be made available for uses that accord with these principles.

Google also commits to not pursuing:

1. Technologies that cause or are likely to cause overall harm.
2. Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
3. Technologies that gather or use information for surveillance, violating internationally accepted norms.
4. Technologies whose purpose contravenes widely accepted principles of international law and human rights.

But there are few specifics on how it will actually do so.

AI applications can cause a wide range of harms

Google's principles recognise AI's risk of bias and its threat to privacy. This is important in light of the findings that Google search algorithms can reproduce racial and gender stereotypes. But the principles fail to acknowledge the wider risks to all human rights and the need for them to be protected. For example, biased algorithms not only result in discrimination but can also affect access to job opportunities.

Aside from the search engine, Google's other businesses could also raise human rights issues. Google created the company Jigsaw, which uses AI to curate online content in an attempt to address abusive language, threats, and harassment online. But content moderation can also pose

threats to the right to freedom of expression.

Google Brain is using machine learning to predict health outcomes from medical records, and Google Cloud will be collaborating with Fitbit. Both of these examples raise privacy and data protection concerns. Our colleagues have also questioned whether partnerships such as Google DeepMind and the NHS benefit or undermine states' obligations to put in place a healthcare system that "provides equality of opportunity for people to enjoy the highest attainable level of health."

What should Google do?

Google's overall approach should be based on finding ways for AI to be beneficial to society without violating human rights. Explaining its first principle to be "socially beneficial," Google says that it would only "proceed where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides." But an approach that balances the risks against the benefits is not compatible with human rights. A state or business, such as Google, cannot develop an AI that promises to benefit some people at the expense of the human rights of a few or a particular community. Rather, it has to find a way to ensure that AI does not harm human rights.

So, Google needs to fully consider the effects of AI on human rights throughout its development and deployment. Especially so, because risks can arise even if the technology is not designed for harmful purposes. International human rights standards and norms – including the UN Guiding Principles on Business and Human Rights – cover both the purpose and the effect of actions by businesses, including Google, on human rights. These existing responsibilities need to be much more clearly reflected in Google's principles, particularly on the positive action that Google will take to protect harm to human rights, even if unintentional.

To be responsible over how it develops and deploys AI, Google needs to move beyond the current tentative language about encouraging architectures of privacy and ensuring "appropriate human direction and control" without explaining who decides what is appropriate and on what basis. It needs to embed human rights into the design of AI and incorporate safeguards such as human rights impact assessments and independent oversight and review processes into the principles.

The principles should also detail how harms to human rights will be remedied and how individuals and groups affected can bring a claim, which is currently absent.

The way forward?

Launching the principles, Google's CEO Sundar Pichar recognised that the way in which AI is developed and used will have "a significant impact on society for many years to come." Google's pioneering role in AI means that the company, according to Sundar, "feel[s] a deep responsibility to get this right."

Although the principles are an important start, they need much more development if we are to be assured that our human rights will be protected. The next step is for Google to embed human rights, safeguards and accountability processes throughout their AI development. That is what is needed to "get this right."

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: Google's new principles on AI need to be better at protecting human rights (2018, June